DOCUMENT RESUME

ED 069 711 TM 002 163

TITLE	Constructing and Using Achievement Tests: A Guide for Navy Instructors.
INSTITUTION	Department of the Navy, Washington, D.C. Bureau of Naval Personnel.
REPORT NO	NAVPERS-16808-B
PUB DATE	71
NOTE	109p.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (GPO 433-578/20)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Achievement Tests; Essay Tests; Grading; Identification Tests; *Manuals; *Performance Tests; Scoring; Teaching Guides; *Test Construction; Testing; Test Interpretation; Weighted Scores

ABSTRACT

This revised manual has been designed to be used by Navy instructors in shore-based schools, training afloat, and components of the Naval Reserve as a guide for the construction and use of achievement tests. The seven chapters cover: 1. Navy Training and Achievement Testing; 2. Performance and Identification Tests; 3. Written Tests; 4. Administration of Tests; 5. Scoring Tests and Grading Students; 6. Interpretation of Test Results; and 7. Weighting and Combining Test Scores. (JS)

ED 069711

CONSTRUCTING AND USING ACHIEVEMENT TESTS

A Guide for Navy Instructors

NAVPERS 16808-B

1

ED 069711

CONSTRUCTING AND USING ACHIEVEMENT TESTS

A Guide for Navy Instructors

PREPARED BY
BUREAU OF NAVAL PERSONNEL

NAVPERS 16808-B

UNITED STATES
GOVERNMENT PRINTING OFFICE
WASHINGTON

PREFACE

The first edition of *Constructing and Using Achievements Tests* was published in 1945 for the primary purpose of assisting instructors and training officers attached to service schools in preparing progress and final achievement examinations in their courses. Copies of the manual were distributed to leading authorities in the field of educational measurement for comment and criticism. The response to this circulation was universally favorable and produced a number of suggestions for improvement in the organization and content of the manual. The purpose of providing a working guide rather than a theoretical treatise on test development was recognized and endorsed by these experts.

The present revision, prepared in the Bureau of Naval Personnel, incorporates many of the suggestions made by the measurement experts mentioned above, includes the material on performance and identification testing within the manual itself, expands the chapter on utilization of test results, and adds a chapter on the combining and weighting of test scores.

In its revised form this manual should be used by every instructor in shore-based schools, training afloat, and components of the Naval Reserve as a guide for the construction and use of achievement tests. Only when it is used by those most directly concerned with the training of naval personnel will it be making its maximum contribution to the Navy's training program.

TABLE OF CONTENTS

CHAPTER I. NAVY TRAINING AND ACHIEVEMENT TESTING

Section 1A. Navy instruction 1
1B. Navy tests 2
1C. Kinds of tests 3

CHAPTER II. PERFORMANCE AND IDENTIFICATION TESTS

Section 2A. Typical jobs for the performance test 5
2B. Some basic rules 5
2C. A most common error 7
2D. The wrong way 8
2E. The right way 8
Illustration I — 20MM performance test 10
Illustration II — Circuit tracing and hook-up performance test 14
Illustration III — Torpedo performance test 20
2F. Why use identification tests? 23
2G. Kinds of identification tests 23
2H. Constructing the identification test 24
2I. How it is done 24

CHAPTER III. WRITTEN TESTS

Section 3A. Planning the test 28
3B. Types of questions 29
3C. Writing the multiple choice item 29
3D. Matching items 37
3E. Completion items 42
3F. True-false items 43
3G. Essay questions 44
3H. How long should the test be? 46
3I. Check list for building the written achievement test 46

CHAPTER IV. ADMINISTRATION OF TESTS

Section 4A. A general rule 49
4B. Specific points for the written test 50
4C. Coordination of identification, performance, and written tests 52

CHAPTER V SCORING TESTS AND GRADING STUDENTS

Section 5A. Mechanism of scoring 56
5B. Use of the separate answer sheet 57
5C. Converting test scores into student grades 60

TABLE OF CONTENTS

(continued)

CHAPTER VI. INTERPRETATION OF TEST RESULTS

Section 6A. Getting full value from testing 75

6B. Analyzing areas of failure 75

6C. Making better tests 79

CHAPTER VII. WEIGHTING AND COMBINING TEST SCORES

Section 7A. The need for weighting scores 82

7B. Procedure for weighting 86

7C. Suggestions on the use of a weighting system 92

CHAPTER 1

NAVY TRAINING AND ACHIEVEMENT TESTING

1A. NAVY INSTRUCTION

Maintaining the fleet in strength and readiness, the Navy's major peace-time responsibility, is largely a matter of the training of men. The instruction needed for the training of men is a continuing function of the training program. However, there are two general phases of the training program: (1) school instruction and (2) "on-the-job" training.

Instruction given at Navy schools trains in the basic ideas and understandings needed for proper performance of the various Navy jobs which the men will be called upon to do. The instructional program at a Navy school is a concentrated program aimed toward the objectives of teaching fundamental knowledge and developing skills on the part of each trainee so that he will be better prepared to do the job he has to do in maintaining the fleet in strength and readiness.

"On-the-job" Navy training provides a man with the opportunity to broaden his knowledge and further develop his skills and abilities so that he is able to do better the work he is doing. A second purpose of "on-the-job" training is to prepare a man to assume greater responsibilities and enable him to advance in his Navy rating.

How well the Navy does its job of training depends greatly upon the skill and efforts of the Navy instructors who handle the daily task of getting the "know-how" into the man.

Every phase of the work of these Navy instructors calls for a high order of technical skill. Nowhere is this more true than in the testing of a man in training to determine how much of that all-important "know-how" he has learned. Any experienced teacher will testify that it is often harder to ask a good question than to answer one; that it is a bigger headache to make a test than to take one. Yet a good test serves so many significant purposes in the training program that most instructors will want to improve their skill in the measurement of student achievement.

Testing is one of the important phases of any training program. It would be a foolish procedure to attempt to administer a training program without some effort to measure the effectiveness of the training. Everyone gets tested somehow. If not by a formal testing program carried out while the training is being given, then cer-

tainly by formal or informal evaluation of how well he performs his duties on the job after training is completed. It is more efficient to have a continuing check on a man's progress during his period of training, to determine his qualifications for doing the job, than to wait until *after* he has been assigned to the job to see whether or not he has the knowledges and skills which his billet will demand. In addition to being a short-cut to evaluate a man's ability to perform on the job, organized testing serves as a safety measure to prevent future loss due to unskilled performance on the job itself.

Testing is the evaluation of performance in relation to a standard, whatever that standard may be. It has been shown that testing, properly carried out, gives a more accurate measure of a man's qualifications to fill a billet than can be obtained by the personal judgment of one individual or even a group of individuals under the usual conditions encountered in the Navy.

This manual is designed to offer technical assistance to (1) instructors in schools and service commands in their efforts to construct and use adequate tests for their trainees, and (2) officers and petty officers concerned with local training programs to help them prepare sound texts to evaluate their programs. It answers three basic questions: (1) Why test? (2) How are the various types of tests constructed? (3) How should test scores be interpreted and used in assigning grades?

1B. NAVY TESTS

In the Navy tests are used for a variety of purposes. Some tests have been developed for use in selecting men for particular schools or billets. Other tests are designed to classify men as to their various abilities. A third type of test is used to measure the progress and achievement of men in schools and in training programs. Still other tests are used to determine a man's eligibility for advancement or promotion. These third and fourth types of tests are classified as achievement tests. It is with the construction and use of these achievement tests that this manual is mainly concerned.

Achievement tests are used in each of the three phases of instruction: (1) as pretests,—(2) as progress tests, and (3) as qualifying tests.

1. The pretest.

When used as a pretest, an achievement test helps determine the status of the previous knowledge of the trainee before the instructional program is begun. A test to be used at the beginning of a training period should be designed to cover the whole job for which the man is being trained. It should be comprehensive in scope, but general in application. The results of a pretest indicate the level towards which future instruction should be aimed. The test can also select those men who can profit immediately from more advanced training than is given in the instruction covered in the course. It is a waste of the Navy's time and money to give instruction to a group of men on materials these men already know.

It is equally unwise to give advanced instruction to men who do not know the basic fundamentals of a subject.

2. The progress test.

A progress test used during the period of instruction helps to give an answer to the questions, "How well has this material been taught?" "What parts of the instruction need reteaching?" and "Is the rate of progress satisfactory?" Just as a bricklayer constantly uses his plumb bob and level to check his progress in the laying of a brick wall, so should a Navy instructor use achievement tests to measure the progress of his teaching and of his men's learning throughout the entire instructional program.

3. The qualifying test.

Used at the end of the period of instruction, an achievement test serves as a qualifying test. It helps to give the answers to the questions "Is this man now ready to do the job for which he has been trained?" and "Has he mastered sufficiently the skills and knowledges needed to insure success in his next assignment?" An end-of-course test should be broad in scope and detailed in coverage.

An achievement test is a means for measuring what the student has learned. A good achievement test can help tell you:

a. What the student knows; what he can do. This will help:
   (1) Determine if he is ready to advance to the next lesson, the next rate, another school, or to the fleet.
   (2) Encourage and challenge the trainee. A good test program keeps a man on his toes. It tells him where he's strong, where he's weak, where he stands.
   (3) Make accurate and fair grading possible.

b. What the student doesn't know; what he cannot do. This will:
   (1) Enable the instructor to fill in important gaps in the student's knowledge and select students needing special attention. A test used this way serves a far more significant purpose than a test used only to get grades for the men.
   (2) Point the way to the correction of teaching weakness. A test is a trial of both students and instructors. Gaps in teaching efficiency require the same caulking as gaps in the student's knowledge.
   (3) Eliminate the "duds" and "gold-bricking" students.

1C. KINDS OF TESTS

There are many different kinds of tests, but the types most frequently used in the Navy are *performance* tests, *identification* tests, and *written* tests.

1. Performance tests.

In these tests the student is required to *do* all or some part of the job for which he is being trained. This may involve repair, assembly, disassembly, or operating

procedure. To build this type of test, the instructor must first decide on the essential skills and information a man must have to do the job for which he is being trained. Selected tasks will then be used to determine if the man has reached the necessary level of proficiency in performance. While such a test works well to measure skill of a gunner's mate in the assembly or disassembly of machine guns, or the ability of a radioman to send or receive messages, it is not so acceptable as a means of testing knowledge of theory or principles. To test for information on what takes place in a vacuum tube or how various types of guns are cooled, another kind of test would be used—the written test.

2. Identification tests.

These tests give a realistic appraisal of what the trainee learned during the training period and stimulate the trainee to become familiar with Navy equipment. Many jobs in the Navy require that a man be thoroughly familiar with the parts of equipment and the functions of these parts in order that he correctly perform his duties. "I want the 'gizmo' that fits on the 'what-ja-call-it' that turns the 'thing-a-ma-jig' around the 'whoosit' " may be meaningful to a machinist's mate as he asks for a specific spur gear, but the storekeeper in charge of the tool crib is left completely in the dark. There are just too many different pieces of technical equipment to permit their classification as "what-nots" and "do-hickeys." Identification tests used early in the training program determine a man's ability to recognize a piece of equipment, a picture of that equipment or a verbal description of it, and his knowledge of the function or use of that part.

3. Written tests.

Many different kinds of written questions can be developed to measure a student's knowledge of a subject. Such tests are generally used to determine a student's *information about* and *understanding of* principles and practices. Such information and understanding are generally present when skill in performance is present. However, a written test can often aid in predicting how well a man will actually perform in a practical operating situation. It may often prove valuable to use both a written and a performance test to secure a more complete judgment of a man's over-all proficiency.

CHAPTER II

PERFORMANCE AND IDENTIFICATION TESTS

2A. TYPICAL JOBS FOR THE PERFORMANCE TEST

A great many of the jobs in the Navy require development of skill in performing manual operations or in dealing with specific materials, tools, or machines. These are situations ideally suited for testing proficiency by samples of performance. Some typical operations which can be checked by the performance-type test follow:

1. Typing a letter
2. Assembling a particular gun
3. Sending a semaphore message
4. Charging a storage battery
5. Tuning a radio transmitter
6. Replacing carbons in a searchlight
7. Preparing foods
8. Splicing a wire cable
9. Entering data into the computing machine of a fire control system
10. Removing a torpedo detonator
11. Finding the range of a target
12. Locating a radio short circuit
13. Wiring a starting box to a shunt motor
14. Sewing parachutes
15. Caulking a seam
16. Assembling aerial photographic mosaics

The range of situations in which performance tests can be used is very broad, much broader even than the varied situations above indicate. This list is merely suggestive of a few of the many places where performance tests may serve as a measure of the essential skills the Navy tries to develop.

2B. SOME BASIC RULES

There is nothing hard and fast about the way to build a performance test. However, for most situations there are certain basic steps that must be followed in order to avoid getting fouled up. Getting the answers to the following questions in the

order in which they are presented will aid in the development of a logical and sound test.

Question 1: *What are the essential skills and information these men must have to do the job for which they are being trained?* This question goes far beyond what they have read, what lectures they have heard, what demonstrations they have seen, what work they have done. Every part of the test must be examined in terms of whether it gets at an essential skill or knowledge related to the job a man will do when his training is completed. Whether some particular thing was taught, whether some subject was treated in a text or a movie or a demonstration is only important as it relates to the main purpose of the training which a man undergoes. If non-essential material is taught, or subjects of minor significance are dealt with, there is no justification for making these the basis for a test. The major question is always, "What skills must this man possess as a result of this training?" It is not, "What has this man been taught?"

Question 2: *What tasks best represent the most significant aspects of the performance you want to test?* Frequently it will be necessary to use performance in a part of the job to tell you how well the man can do the whole job. In testing men on the disassembly of a torpedo, it would be impractical to have each man do a complete job. The use of enough representative tasks, selected for their coverage of the most significant phases of performance, will provide a good measure of the achievement of each man.

Question 3: *What features of performance on the tasks selected will be used to indicate good or poor performance?* Having chosen the particular tasks which are most representative and significant, it is now necessary to determine the standards a man is expected to meet in carrying out these tasks. This means, of course, reaching a concrete decision on the particular elements in performance which are considered important and which will be marked or scored. There are a number of such elements which may be considered with reference to the performance of any task. Some of the most useful measurements are:

a. *The quality of the finished job.* Example: Men are given the proper tools and materials and told to splice the ends of a cable. No observation is made of the procedure used. Performance is judged on a number of specific points related to the finished product, for example surface appearance, cleanness of joint, conformance to specified tolerance, etc.

b. *The skill and accuracy of operations.* Example: Men are told to solder a broken connection in a circuit. They are judged on the tools and materials selected for the job and on specific points in the procedure they follow.

c. *Speed.* Example: Men are judged on the time taken to rig up a boatswain's chair for repairing a ship's mast.

d. *Combinations.* The scoring elements listed above may be combined in many ways. Quality of the finished job, speed, and skill in operations may all be judged in a single performance test. Which element or elements to consider in scoring will depend on which factors in the performance being tested are considered most important.

Question 4: *What is the most effective, consistent, and economical way to give the test?* This may involve determining:

a. How long it will take a given group of students to complete all the tasks.

b. What equipment will be needed.

c. How the equipment can be organized in order to insure a smooth, steady flow of students from one task to another.

d. What assistance will be needed in giving the test.

e. What specific directions will be needed for the students and for such assistants as timekeepers and scorers. The students must be given a clear idea of what to do and what aspects of their performance will be considered important. If speed is to be scored, the directions to the students should indicate this fact. If the quality of the end product is the only basis for scoring, then this should be stated. Otherwise the students may mistakenly sacrifice accuracy for speed when speed or lack of it plays no part in the test.

f. What record forms, score sheets, check lists, or observation sheets will be required to establish a uniform, clear-cut basis for scoring every student in the same way on every task.

The problems under Question 4 relate to administering the test. Here are two suggestions that will help solve these problems.

(1) Try out the test on a few advanced students or some fellow instructors. If this is not feasible, take it yourself. Valuable information can be obtained as to the time required to give the test, the amount of equipment needed, the best organization of the equipment, and the number of assistants needed.

(2) In addition, prepare a shop layout plan showing where equipment is to be stationed, what equipment will be placed at the various stations, and what tasks will be done at these stations. Planning will pay dividends.

2C. A MOST COMMON ERROR

Suppose you want to find out how fast your men run the 100 yard dash. That's a pretty typical performance test situation. You certainly wouldn't give one man sneakers, another man track shoes, a third man hobnailed boots, and a fourth galoshes. You wouldn't ask them to run without telling them in advance how far to go. Nor would you time the runners with an alarm clock. Instead you would be

careful to give every man the same equipment and clear instructions on what he should do. You would measure performance with as fine an instrument as you could find. Every man who ran that course would get the same task in the same way. You would time each man on the same instrument or on instruments which were proven equal by calibration.

This careful procedure is followed in order to insure that your test is fair to all men and able to make accurate comparisons between men. The same care used in a track meet is equally necessary in the classroom. Yet this is precisely the point at which many performance tests fall down. Tools and equipment often vary from station to station. Instructions vary from class to class. Marking varies from scorer to scorer. While it is true that classroom tests do not generally approach the accuracy of measurement of the 100 yard dash, it is also true that care and forethought will bring a nearer approximation of that accuracy.

Here are two good rules to follow:

(1) ALWAYS GIVE THE TEST TO ALL GROUPS IN EXACTLY THE SAME WAY.

(2) ALWAYS SCORE EVERY STUDENT ON THE SAME BASIS.

2D. THE WRONG WAY

It should be quite clear by now that development of a performance test, indeed any test, requires planning and careful preparation. It cannot be pulled out of a hat even by the most experienced man. Yet it happens not infrequently that a performance test is dreamed up with all the speed and dash of a short order cook scrambling eggs during rush hour. Figure 1 on the opposite page presents a step-by-step analysis of a poorly planned performance test in the field of gunnery. While this subject matter may not apply to your own field of teaching, the illustration will serve to indicate some characteristic pitfalls which may trap the unwary, and will point up the significance of many of the rules for test construction discussed in the preceding paragraphs.

2E. THE RIGHT WAY

There is an almost endless variety of skills that can be measured by performance tests. No single example of such a test can serve as a perfect model. However, a study of the elements in one good performance test can contribute considerably to an understanding of the basic principles common to the construction, administration, and scoring of all performance tests. The descriptions of the three performance tests given on the next few pages are illustrative of the kinds of performance tests which may be developed in almost every type of naval training school.

ANALYSIS OF A POOR TEST ON THE 20 MM GUN

A Poorly Planned Test on the 20 MM Gun

1. "I'd kind of like to know how much you fellows got on that 20 MM gun you studied last week. Let's see. I guess I can pick out a couple of things for you to do to show if you were on the ball. Let me see . . .
2. "Ah yes! Let's use these two guns up front. Suppose you come up two at a time and disassemble the trigger and interlock mechanism on these guns and then you can put them back again.
3. "Don't worry about how you'll be marked. Just do the job and you'll get what you deserve. If you're on your toes, you'll come out all right. If you don't know a trigger from the rear end of a mule, you'll get quite a kick out of your grade. I'll be up here watching every one of you.
4. "I think we can finish this test in an hour if you don't drag your tails. Anyway that's all the time we have. I guess these two guns are enough. Say, maybe we'd better put these guns on the tables in the rear so you men not working on the test can do some work on this model up front. I want the men not taking the test to keep working until they are called. Johnson, lend a hand here. And I guess maybe you'd better help me during the test. I'm not sure how this thing will work.
5. "Now you fellows all know what to do. And Johnson, suppose you mark the men who work on this second gun. Just use your noodle on this thing and I'll help you as we go along. Come on fellows, we've lost fifteen minutes already and we've got to finish today. Let's get going."

Gets An Instructor Into Difficulties

1. With the class in front of you waiting for a test, it's kind of late to start thinking about the skills you want to check. Well, maybe you'll think fast and hit on something important anyway.
2. Just two little jobs to test the most significant skills on this gun? Hardly enough! You can't tell much about a barrel of apples by testing just two little ones off the top.
3. Might as well get the grades right out of the dream book. Measuring performance on this basis is like measuring a yard with a rubber band. What counts in the score here? Speed? Accuracy? Identification of parts; i.e., knowing a trigger from the rear end of a mule? If anyone knows they're keeping it a secret.
4. Heading for the rocks. No clear idea as to whether the equipment is adequate to do the job in the time available. Time lost and confusion present in setting up equipment, all of which should have been arranged at fixed stations in advance of the test period. Only a guess on the number of assistants needed. You wouldn't organize battle stations so haphazardly. Your strategy (knowing what your aim is) and your tactics (knowing how to achieve that aim) must be carefully planned in advance.
5. The students know they have two tasks to do, but they don't have a cloudy idea as to what will be looked for when they do them. Shall they work fast and risk errors? Or slowly and accurately? As for the assistant, the ceiling is zero with more bad weather ahead.

Because These Questions Were Not Raised and Answered

1. What major skills were to be developed by studying this gun? Repair of the gun? Or the firing of the gun? Or assembly and disassembly? This is the first problem to be solved.
2. What are the most significant tasks related to these skills which indicate whether the necessary proficiency was developed? Choose *enough representative* tasks to really test the men.
3. What will be the standards by which you will judge good or poor performance? In scoring, will you judge speed of performance? Or the quality of the end product? Or some other factors? Are these features the important aspects of the performance which will govern good or poor work when the men go on the job?
4. How can the administration of the test be organized most effectively? Make a shop layout plan and try the test ahead of time to determine (a) how much time it will take to give the test, (b) what equipment will be needed, (c) where equipment should be stationed, and (d) how many assistants will be needed. The trial of the test should also help you check on the effectiveness of the standards you have set for scoring. Organize the test so every man is busy for the whole period.
5. Do students, timekeepers, scorers know exactly what to do and how to do it? Do the students know on what basis they will be scored? Do the scorers know just how to score? Do they have standard score sheets? Will everyone be scored in the same way on the same task?

Figure 1.

ILLUSTRATION 1:

A 20 MM PERFORMANCE TEST

PLANNING SHEET

I. **Purpose of the test:** Study of this gun is designed primarily to give the men facility in the disassembly and assembly of various parts and mechanisms. Speed and accuracy of operations are both considered important. Ability to identify various parts is also required. However, since a separate test has been prepared covering identification of parts, this performance test deals mainly with disassembly and assembly operations. Recognition of parts by name is incidental to this test and should not enter directly into the scoring.

II. **Items (Tasks):**

a. Disassembly—assembly of the trigger mechanism.
b. Disassembly—assembly of the magazine interlock mechanism.
c. Disassembly—assembly of the double-load stop mechanism.
d. Removal and replacement of the barrel.
e. Removal and replacement of the breech face piece.
f. Cocking and uncocking the gun.
g. Shipping and unshipping loaded magazine.

III. **Time required per student:** A trial run on the above seven tasks shows that the first three tasks should take no longer than seven minutes. The next four tasks should also be completed within seven minutes. Allowing time for restoring gear, etc., the test time for each student is twenty minutes.

IV. **Equipment and organization of equipment:**

a. For tasks a to c, a 20 MM gun (with shoulder bars removed) on an assembly table is needed.
b. For tasks d to g, a 20 MM gun fully assembled on mounts is needed.
c. Standard tools for assembly—disassembly, magazine loading, and breech face removal are required.
d. With six guns available, six students could be tested in twenty minutes. Eighteen students could be tested in an hour. Fifty-four students could be tested in three hours.
e. Three guns on assembly tables will be set up at Station 1. Each will be used to test students on the first three tasks. Three guns on mounts will be set up at Station 2 for the next four tasks.

V. **Assistance needed:** One assistant at each gun. One timekeeper. Total: 7 men.

VI. **Scheduling:** By running this test in conjunction with a written examination, it will be possible to send men to the shop where the performance test is given in groups of six at intervals of twenty minutes. This will keep a constant flow of men to the test room, prevent observation of the test by men not engaged in taking it, and insure that all students will be kept active throughout the period.

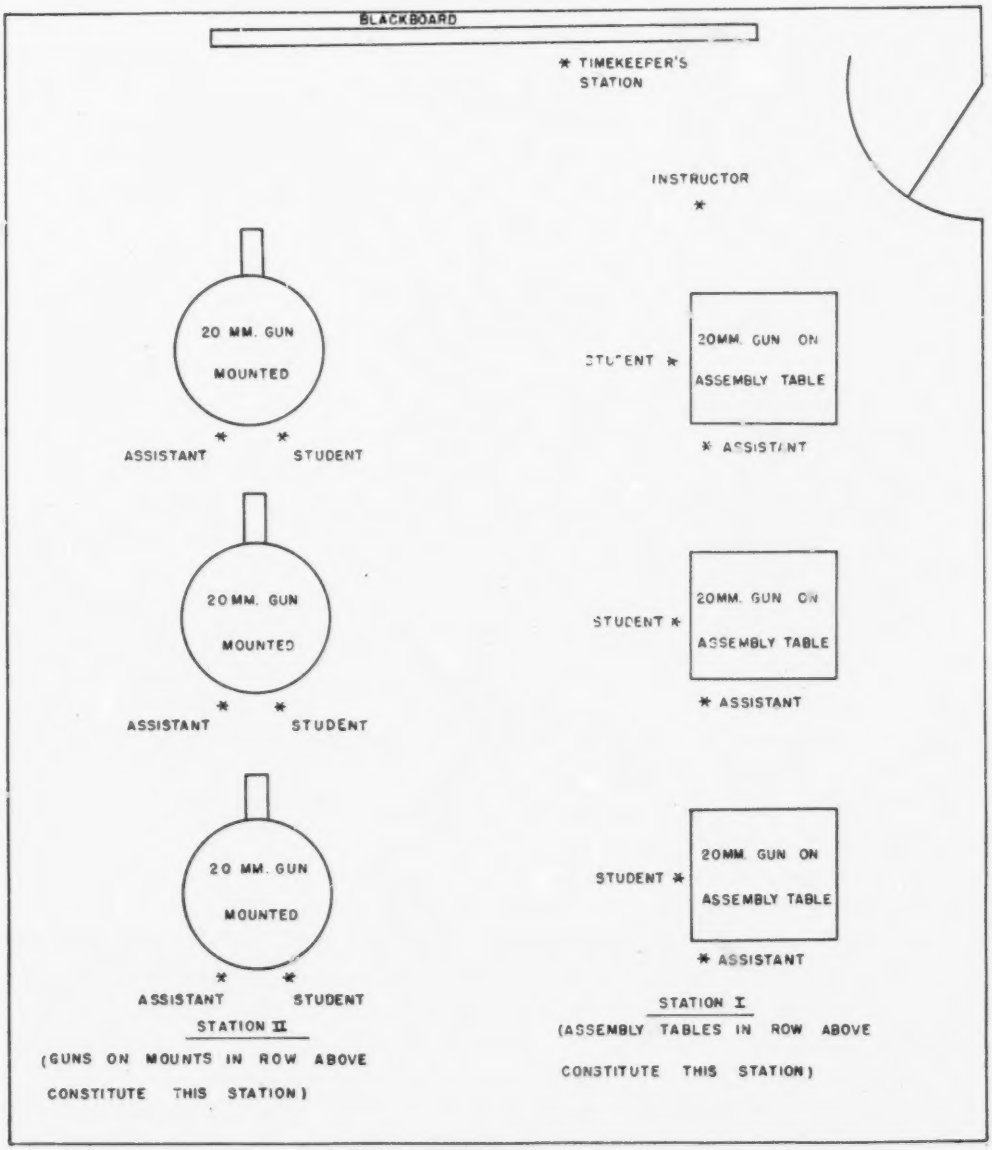


Figure 2.—Shop layout for 20 MM performance test.

20 MM PERFORMANCE TEST
DIRECTIONS SHEET

I. Directions to timekeeper

1. As each group to be tested enters the shop, give each man a Record Sheet and assign him to a gun at one of the stations
2. When students are at stations, read DIRECTIONS TO STUDENTS aloud.
3. After reading these directions, and answering questions, call "READY, GO."
4. Immediately place a big "1" on the blackboard. At the end of the first minute, change to "2," and so on through the 7th minute.
5. At end of 7 minutes call "STOP." Erase "7" from board immediately.
6. When gear is restored by students and assistants, call "CHANGE STATIONS."
7. Repeat above procedure with students at exchanged stations.
8. Collect all Record Sheets after each group has worked at both stations. Dismiss men. Call in next group.

II. Directions to students (to be read by timekeeper.)

1. "You are going to be tested on your ability to *disassemble and assemble various parts of the 20 MM gun*. Here's the procedure.
2. "When I give the order, 'READY, GO,' you will be told by the assistant at your station what to do. Start right in working and work fast. You will be marked on every task for the length of time it takes you to finish that job. Even though time is important, don't work so fast that you get all fouled up. You will have seven full minutes at each station to do all the jobs of that station. Steady, sure work will get you through in good shape.
3. "Stop immediately on the signal 'STOP.' Help the assistant to get all gear restored to its original shape.
4. "Don't move from your station until the order to 'CHANGE STATIONS' is given or you are dismissed.
5. "Are there any questions?"

III. Directions to assistants

1. As each student enters the shop he will be given a Record Sheet by the timekeeper and assigned to a gun at one of the two stations. Take the Record Sheet and record on it the name and class number of the student.
2. Give orders on each task as listed on the Record Sheet, only after timekeeper says "Go". Give orders *quickly* and *plainly*. This is important.
3. Stop student's work immediately at order "STOP". Check scoring and restore gear for the next student to be tested.
4. Scoring. On Record Sheet check each incomplete task. Record time it takes to finish all tasks. This time will be the number you see on the blackboard at the instant the student finishes all work at your station. It will be placed on the blackboard by the timekeeper. If a student fails to complete any task at your station, his time score must be entered as 8.
5. No student may leave your station to go to the next station until the order to "CHANGE STATIONS" is given.
6. Do not interfere with student's performance; do not advise, encourage, or stop student from making errors, unless the error may prove dangerous.

20 MM PERFORMANCE TEST

RECORD SHEET

Name_____ Class_____

Station 1. **Check Incomplete Tasks**

1. Take out the trigger plunger and spacer. _____
2. Reassemble the trigger mechanism _____
3. Remove magazine interlock carrier spring. _____
4. Reassemble the magazine interlock mechanism. _____
5. Take out the DLS lever spring. _____
6. Reassemble the DLS mechanism. _____ Time_____

Station 2.

1. Remove the barrel from the gun. _____
2. Replace the barrel. _____
3. Remove the breech face piece. _____
4. Replace the face piece. _____
5. Cock the gun. _____
6. Uncock the gun. _____
7. Place loaded magazine on the gun. _____
8. Remove loaded magazine from the gun. _____ Time_____

Total Incomplete _____

Total Time_____

Total Score_____

Note on Scoring Procedure

Since each incomplete task counts 1 point and each minute of time counts 1 point, it is possible for a student who doesn't finish any task properly to get a score of 14 for uncompleted work and 16 for elapsed time. His total score would be 30. It should be noted that this is an example where *inverse* scoring is used, i.e., a *high score* indicates that the student has done a *poor* job and should be assigned a *low grade* or *mark* on his performance, and conversely a *low score* indicates *good* performance which should be assigned a *high grade* or *mark*. This fact must be taken into consideration when building a graph or table to convert the scores into grades.

A satisfactory method for converting the test scores to grades or marks is described in this manual in Chapter V, Section 5C.

ILLUSTRATION II:

PERFORMANCE TEST

CIRCUIT TRACING AND HOOK-UP TESTS

PLANNING SHEET

I. **Purpose of the test:** To determine the accuracy with which students can perform the following tasks:

JOB A: To trace circuits by ringing out with a low-voltage buzzer, preparatory to connecting a "Start-Stop" push button switch for operating a magnetic controller. The controller, in turn, connects the motor across the line.

JOB B: To connect a slip-ring induction motor to a magnetic across-the-line starter with push button control.

Speed of performance is of secondary importance. Accuracy of performance is the first essential.

II. **Items (Tasks):**

JOB A:
1. Identification of controller terminals.
2. Identification of leads connecting motor to controller.
3. Identification of lines between push button and controller.
4. Identification of push button terminals.

JOB B:
1. Wiring main line correctly to line terminals of controller.
2. Wiring motor connections for correct voltage (220 to 440).
3. Wiring motor correctly to the three motor terminals of controller.
4. Completion of slip-ring circuit with resistance.
5. Connection of control circuit push button to proper terminals.

III. Time required per student: A trial run on the tasks involved in Jobs A and B indicates that each student should take no longer than twelve minutes for each job. Allowing time for restoring gear, checking scores, etc., the test time for each student on both jobs will be thirty minutes. With equipment available, six students can be tested at one time. Twelve students can be tested each hour.

IV. Equipment and organization of equipment:

JOB A:
1. Three "Start-Stop" push button switches mounted on wall panel. Circuit diagrams posted above panel.
2. Three magnetic controllers. All controller terminals are led to terminal posts on a panel at the rear of the controllers. Circuit diagrams should be posted on the inside panel of each controller door.
3. Tools: Crescent wrench and leads placed on top of each controller. Portable buzzers with leads attached. Chalk for marking terminals.

JOB B:
1. Three slip-ring induction motors, magnetic across-the-line starters with push button controls, and external three-phase resistance.
2. Tools: Screw drivers, crescent wrenches and leads placed on table beside the motors.

Station I is represented by the equipment for Job A. Station II is represented by the equipment for Job B. The students at Stations I and II will exchange places when directed to do so by the timekeeper so that each student has an opportunity to work on both Jobs A and B.

V. Assistance needed:

A. One instructor to act as timekeeper and to supervise and score the students' work on Job A.

B. One instructor to supervise and score students' work on Job B.

VI. Scheduling: These tests are to be run in conjunction with the regular shop work of the students. Six students will be drawn from the shop class at half-hour intervals and given the tests. The remaining men in the class will continue with their regular shop assignments. This will insure that all students are kept active and will prevent observation of the testing procedures by men not engaged in taking the tests.

CIRCUIT TRACING AND HOOK-UP TESTS

DIRECTIONS SHEET

I. Directions for assistant at station I (Job A)

1. This assistant supervises test, acts as timekeeper, and scores students at Station I.
2. Before test begins, opens controller doors, removes covers from push button boxes.
3. Prepares scoring key indicating the correct markings of all terminals on the magnetic controllers and related push button boxes. Key is used to check work of students.
4. When each group of students reports for the test, assigns three to equipment at Station I and three to equipment at Station II. Each student should be given a Record Sheet to fill in with name and class. Collects Record Sheets for students at Station I.
5. Reads "DIRECTIONS TO STUDENTS" aloud. Answers questions. Then says "READY, GO."
6. Immediately places "1" on blackboard. At beginning of second minute, changes to "2." Continues through twelfth minute. At the end of the twelfth minute calls "STOP."
7. Checks and scores each student at Station I. Each terminal or lead correctly marked is given a credit of 1. A "1' should be placed in the appropriate boxes provided on the Record Sheet. Enters the total in the box labelled "JOB A: Total Correct."
8. Sees that all gear is restored to original place and all chalk marks thoroughly erased.
9. Directs students to "EXCHANGE STATIONS." Exchanges Record Sheets with instructor at Station II. Repeats procedure in paragraphs 5 through 8 above, except that in reading "DIRECTIONS TO STUDENTS," only paragraphs 2 and 3 need be repeated.

II. Directions for assistant at station II (Job B)

1. Collects Record Sheets after each student has written in his name and class.
2. At the end of each twelve minute test period checks and scores each student's work. Enters a "3" in the appropriate box on the Record Sheet for each item done correctly. Enters the total credits earned in the box labelled "JOB B: Total Correct."
3. After recording scores sees that all gear is restored to original position.

III. Directions to students (To be read by assistant at Station I).

1. "You men are going to be tested on your ability to do two jobs. One job is to *trace out circuits with a low-voltage buzzer.* The other job is to *connect a slip-ring induction motor to a 'magnetic across-the-line starter'* with push button control. Here is the way we are going to work.

2. "The three men along the wall who have been assigned to these magnetic controllers are to use the low-voltage buzzer and ring out all the controller terminals and all push button terminals on the wall panel. Mark each terminal in accordance with the lettering and numbering on the circuit diagrams. Use chalk to mark each terminal. Remember, your job is to ring out and mark every terminal. Your score will be the number of terminals correctly marked, so be accurate and careful.

3. "The three men who have been assigned to the 'across-the-line starters' and slip-ring induction motors on the assembly table are required to make all connections between the motors and starters. Your score will be the number of correct connections made.

4. "Each man is to work on his own equipment. You are not to begin work until the signal to start is given. You will have twelve minutes for each of the two jobs. Steady work without hurry will get you through in good time. If you finish before the twelve minutes are up, check your work and stand by. Do not move from your station or examine your neighbor's work.

5. "At the end of the twelve minutes the signal to stop will be given. Wait until your work is checked and scored and then restore all gear to its original position. If you have been identifying terminals, be sure that all chalk marks are thoroughly erased. When you are told to 'exchange stations,' the men working on the identification of terminals and leads will change places with the men connecting the across-the-line starters to the motors. Are there any questions?"

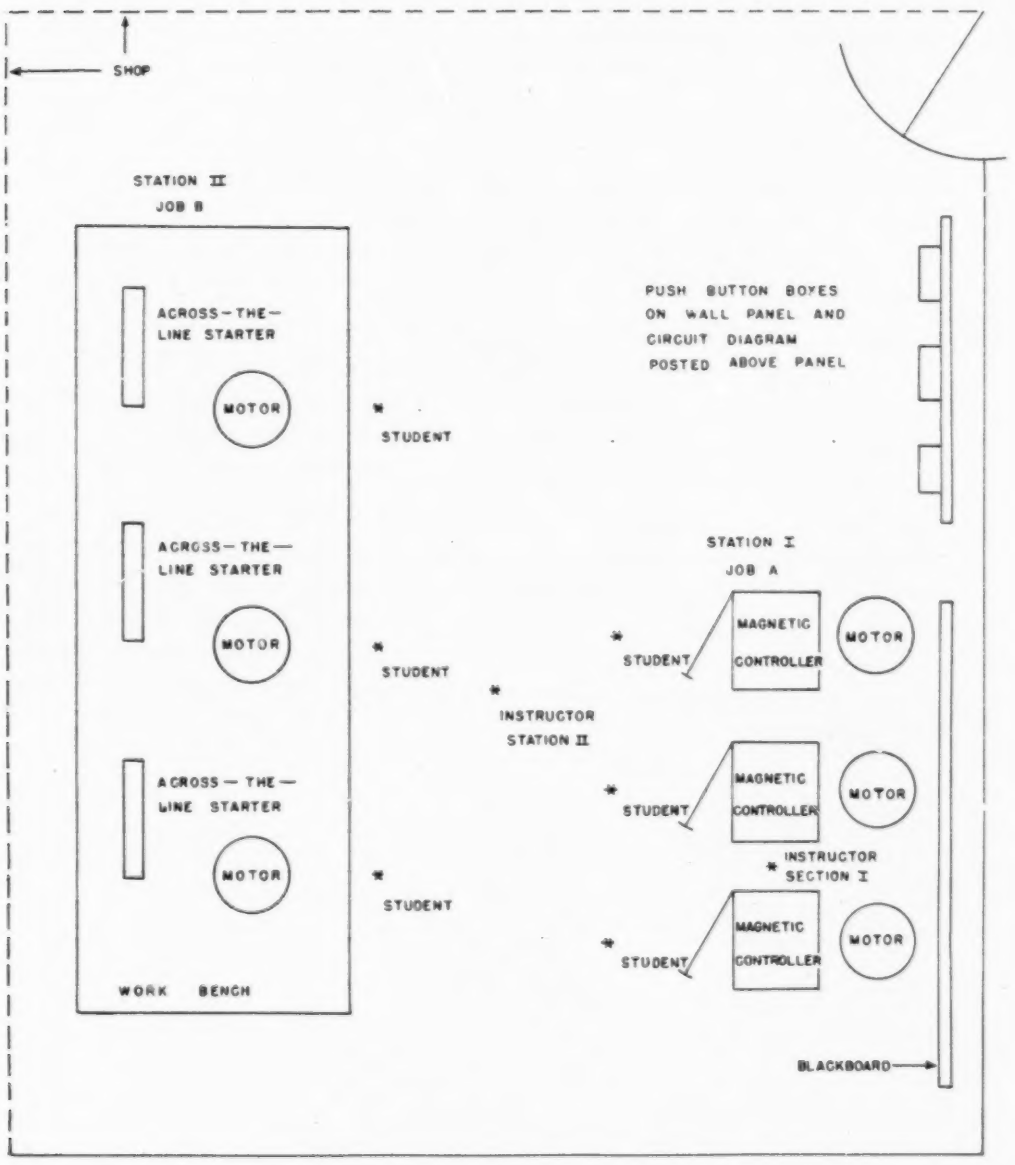


Figure 3.—Shop layout plan circuit tracing and hook-up test.

CIRCUIT TRACING AND HOOK-UP TESTS
RECORD SHEET

NAME_____ CLASS_____SCORE ☐

Station 1: Job A.

To trace out circuits and prepare to connect a "Start-Stop" push button switch to operate a magnetic controller which connects the motor across the line.

		ONE POINT FOR EACH TERMINAL
I. Identification of Controller Terminals	P1	☐
	P2	☐
	P3	☐
	P4	☐
II. Identification of Leads Connecting Motor to Controller	A	☐
	B	☐
	C	☐
III. Identification of Lines between Push Button and Controller	1	☐
	2	☐
	3	☐
	4	☐
IV. Identification of Push Button Terminals	P1	☐
	P2	☐
	P3	☐
	P4	☐

JOB A: Total Correct ☐

Station 2: Job B.

To connect a slip-ring induction motor to a remote, push button controlled, magnetic across-the-line starter.

	THREE POINTS FOR EACH ITEM
I. Main line wired correctly to line terminals of controller	☐
II. Motor connections wired for correct voltage (220 or 440)	☐
III. Motor connected correctly to three motor terminals of controller	☐
IV. Slip Ring Circuit completed with resistance	☐
V. Control Circuit Push Button connected to proper terminals	☐

JOB B: Total Correct ☐

JOBS A AND B: Total Correct ☐

The raw score may be converted to a grade or mark in the manner described in this manual in Chapter V, Section 5C.

ILLUSTRATION III

A TORPEDO PERFORMANCE TEST

PLANNING SHEET

I. **Purpose of the test:** To determine the accuracy with which students can enter the depth steering line and calibrate the depth spring in the depth engine of the Mark 13 or Mark 15 torpedo.

II. **Points of rating (Items):** The steps in the two processes involved in this test are listed on the Record Sheet. Each student's performance is rated on each of these steps.

III. **Time required per student:** Ten minutes should be ample time for most students to complete all the required steps. Approximately two minutes are needed for the proctor to ready the gear for the next student.

IV. **Equipment and organization of the equipment:** The equipment needed for this test includes:

a. A depth engine of either the Mark 13 or the Mark 15 torpedo, set up with housing on leveling stand, air connected to the engine, transportation pin out, hook installed in lower spring socket, and scribe marks out of line.

b. Tools: 11, 48, 49, 180, 246, 411B, WE3.

One complete set-up is required for each student being tested. As many students can be tested simultaneously as there are set-ups available. By running this test in conjunction with other similar torpedo performance tests, several students may be tested at the same time, each student working on a different test and rotating to other tasks at the end of each test, until all students have performed the tasks involved in the several tests.

V. **Assistance needed:** One instructor acts as timekeeper and supervises the general administration of the test. One proctor is required for each student.

VI. **Scheduling:** This test may be run in conjunction with the regular shop work of a class. As many students as can be accommodated by the test equipment may be drawn from the shop class for the required time and then returned to their regular work, being replaced at the testing stations by another group. This procedure will insure that all students are kept active and will prevent observation of the testing process by men not engaged in taking the test.

DIRECTIONS SHEET

I. Directions for test supervisor (Timekeeper).

1. This instructor supervises the entire test and acts as timekeeper.
2. Before the test begins, has proctors check to see that the equipment is in readiness for the trainees. Issues Record Sheets to proctors.
3. When each group of students reports for the test, assigns one man to each of the test stations. Has proctors fill in the identification data on the Record Sheet.
4. Reads "DIRECTIONS TO STUDENTS" aloud. Answers any questions. Then says "READY, BEGIN."
5. Notes the time at which trainees begin work. Allows ten minutes for the test. At the end of ten minutes says, "STOP WORKING."
6. Has proctors place equipment in readiness for the next group. Instructs trainees to help the proctors at their respective stations. Allows two minutes for this step.
7. Sends the group to its next assignment.

II. Directions for proctors.

1. Before the trainees arrive, see that the gear is in readiness. The depth engine should be set up with housing on leveling stand, air connected to the engine, transportation pin placed beside tools, hook installed in lower spring socket, and scribe marks out of line. The tools provided are 11, 48, 49, 180, 246, 411B, and WE3.
2. When the trainee reports to your station, fill in the blanks at the top of the Record Sheet. Do not permit him to read the "Steps" listed on the sheet.
3. As trainee proceeds through the steps of the test, record his performance on the Record Sheet. If he does the step correctly, circle the number in the score column (e.g. ①). If the step is wrong or omitted, cross out the number, (e.g. ✗); correct the student before he goes on to the next step. Mark each step by a circle or a cross out.
4. Do not give helping hints to the trainee. If prompting is necessary to prevent injury to equipment or to the man, the step should be counted as wrong.
5. When time is called by the supervisor, restore gear to its original condition. Remove weight, close air valve, displace knurled nut, tilt leveling stand, and leave transportation pin with tools. Have the trainee help in this operation.

III. Directions to students (To be read aloud by test supervisor.)

1. "You men are going to be tested on your ability to do two jobs. One job is to *center the depth steering line in the depth engine of the torpedo* before you. The second is to *calibrate the depth spring of the engine.* The depth engine has been bench-tested.
2. "Each man is to work on his own equipment. You are not to begin work until the signal to start is given. At the starting signal begin the job of centering the depth steering line. When you finish centering the depth line start to calibrate the depth spring immediately. Do not wait for a signal as none will be given.

3. "You will have ten minutes. This time is ample to finish both tasks without rushing. If you finish before time is up, stay at your station; do not examine your neighbor's work.
4. "Think as you work. Your score will depend upon the number of steps you do correctly and in the proper order.
5. "Are there any questions?" (Answer any legitimate questions.) "READY, BEGIN."
6. (At the end of ten minutes.) "STOP. Help the proctor place the gear back in its original condition so that it will be ready for the next group."

IV. Directions for scoring.

The raw score on this test is the sum of the circled scores. Add up the scores which were circled by the proctor and place this figure in the box in the upper right corner of the Record Sheet. This may be done by the supervisor while the next group is being tested, or after the end of the testing session.

SCORE

RECORD SHEET

CENTERING DEPTH LINE AND CALIBRATING DEPTH SPRING (IN STAND)

Name_____ Class_____ Billet #_____

Proctor_____

Steps		*Score*
Doesn't need coaching on how to center steering line.		2
Inserts transportation pin.		2
Opens air valve correctly first trial.		1
Picks 246 for knurled nut, without trying other tool.		1
Loosens clamp screw *before* trying to turn knurled nut.		1
Lines up knurled nut accurately.		1
Hole in knurled nut left in horizontal plane.		1
Secures clamp screw without harmful pressure.		1
Doesn't need coaching on calibration.		2
Levels housing fore and aft.		1
Hangs 16 lb. weight on screw hook.		1
Removes transportation pin.		2
Doesn't change knurled nut.		2
Selects 180 to turn depth spindle, without trying other tool.		1
Turns spindle right way (or quickly corrects error).		1
Aligns scribe marks accurately.		1
Swings pendulum to see if it stops with scribe marks aligned.		1
Knowledge of tools and procedure: (If guessing, cross out both)	Sure of self.	2
	Hesitant in deciding.	1

2F. WHY USE IDENTIFICATION TESTS?

A type of test closely associated with performance is a test of identification. The identification test usually measures familiarity with the names and functions of equipment, or of the tools used in maintaining equipment. It can also be used to measure a trainee's knowledge of the mechanical relationship of parts. Such knowledge is sometimes basic to the development of the skills which are applied in performance, *but* the knowledge can exist without the skill. A man may recognize a butterfly valve on a carburetor and know that it regulates the amount of air admitted to the fuel-air mixture and still be unable to install the valve or adjust the choke wire to operate it properly. On the other hand, a man with highly developed hand skills and "mechanical sense" may be able to put an engine into proper running condition, but unable to make out an order for replacement of a broken exhaust valve push rod because he does not know the technical name for that part.

2G. KINDS OF IDENTIFICATION TESTS

These tests are used in the following situations:

1. **Identification and function of disassembled parts.**

   In this case a part is removed from the equipment or drawn from stock and placed on the top half of a card. On the lower half of the card are placed two lists of answers to be used in testing the men's knowledge of the part. The first list is headed "PART NAME" and contains five choices from which the students select the correct name of the part displayed. The second list, headed "FUNCTION," gives five possible functions or uses of the part displayed.

2. **Identification and function of parts in an operational assembly.**

   It is often difficult and even undesirable to remove parts from an installation. In many situations trainees in the Navy are not required to disassemble certain types of equipment. They are expected to know, however, the names and functions of the various parts of the equipment which they operate. In this test the parts to be used are spotted by numbered tags and students are given a series of item sheets on which the "part name" and "function" lists are printed under the numbers corresponding to the tags.

3. **Identification and function of both disassembled parts and parts in operational assembly.**

   Sometimes an identification test is desired in which some parts are disassembled and others are not. In this situation the test consists of a combination of the first two types described above. For practical reasons, the test stations for the disassembled parts are set up near the operational equipment.

4. **Identification and function of parts shown in a pictorial representation.**

   Often it is advisable to use a picture or a drawing of the part to be identified when such a picture or drawing will serve as satisfactorily as the actual part. It is advisable to use pictures when the part is unusually large and difficulties would be experienced in arranging the part at a test station.

5. Identification and function of parts presented in a verbal description.

In some rare instances verbal descriptions may be used instead of the actual part or a pictorial representation. Extreme caution must be used in developing the verbal description so that it gives no clues to the use or function of the part.

2H. CONSTRUCTING THE IDENTIFICATION TEST

When constructing items to be included in an identification test certain questions must be answered before a good test can be developed.

Question 1. *What are the topics on which the men are to be tested?* The answer to this question depends upon the objectives which have been established for the training. The topics included must be those which measure the students' knowledge of the important and more significant pieces of equipment. Items designed to test obscure parts and trivial details should be kept to a minimum. The fundamental criterion to be applied in the selection of topics is, "Is this knowledge essential?" Not, "What topics have been taught?"

Question 2. *How much testing time is available?* The time to be devoted to a given test will be one of the controlling factors in determining the length of the test. For most identification tests a minute is sufficient time for recognition of a piece of equipment, determining its name and function, and selecting the correct answers from a list of choices. For the whole test, therefore, slightly more than a minute per item should be allowed. If an hour is available for the administration of the test, the number of items included should be no more than 55.

Question 3. *Is the test to be used as a short rough measure or a longer more accurate measure of proficiency?* A short test used to give a relatively rough measure should contain items which are of a more general nature, while a longer test composed of items of a more detailed nature will give a more precise measure of the relative proficiencies of the students.

Question 4. *What equipment is available?* It is a wise practice to develop a "pool of items" covering a field for which identification tests are to be used. It is then possible to make a selection of those items which are designed to test available equipment. It may be necessary to reject certain items because equipment is no longer available or is obsolete.

Question 5. *Is the terminology used within the scope of the vocabulary of the students?* Keep the terminology simple and standard. Items chosen for inclusion in a test should be written in the accepted language of the subject matter field and should fall within the range of the level of vocabulary of the trainees. This applies particularly in wording the statements concerning the functions.

2I. HOW IT IS DONE

The description of a valve identification test given on the next few pages illustrates the kind of identification test which may be developed in various types of naval training situations.

VALVE IDENTIFICATION TEST
PLANNING SHEET

I. **Purpose of the test.** This test is designed primarily to measure the trainee's acquaintance with the various kinds of valves that he disassembles and assembles by testing the trainee's knowledge of
(a) the names of the various parts, and (b) how these parts function.
This test uses actual valve parts instead of verbal descriptions or pictures.

II. **Items (Valve parts to be identified).** This test consists of 29 valve parts selected from the following types of valves: gate, globe, check, bottom blowdown, pressure relief, pressure regulating, boiler feed stop and check, safety, steam trap, boiler feedwater regulator, and main steam stop valve. (An odd rather than an even number of items is chosen for the test in order to permit a smooth flow of students from station to station.)

III. **Time required per student.** A trial run on the identification of the 29 selected parts of valves shows that each identification and determination of function requires no longer than 50 seconds. Allowing time for giving directions and collection of papers, a period of a half-hour is needed to administer this test to 29 students. (If the number of students exceeds slightly the number of items included in the test, dummy stations with blank item cards are included in the circuit to accommodate the few extra men. The total testing time will be slightly increased.)

V. **Assistance needed.**

a. Twenty-nine selected valve parts.

b. Twenty-nine cards (one for each valve part) on which has been typed the test item consisting of a list of five names of parts and a list of five functions of parts. Each part and test item card are known as a station. Each card is numbered to indicate the item and station.
Tables arranged in a hollow rectangle. The arrangement is illustrated in Figure 4.

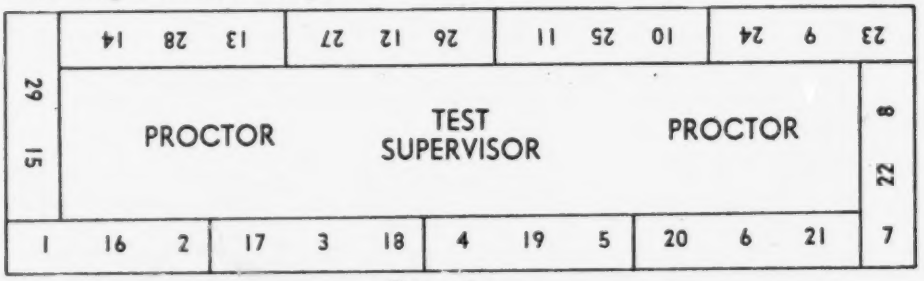


Figure 4

This diagram indicates how the cards (test items) are arranged. The numbers of the two halves of the series are laid out alternately. By moving *two* stations to the right each man answers all items in correct numerical order. By this arrangement, trainees on either side of a man are not working on adjacent parts of the test. This layout reduces the likelihood of mutual assistance, and it also makes it easier for proctors to detect copying. A man must make two complete circuits of the table to complete the test.

d. Separate answer sheets. Answer sheets are needed for each man. Each item has two blanks for recording answers; one blank for the name of the part, the second blank for the function. The answer sheets are laid out at the stations before beginning the test. At each individual station the item number corresponding to the station is circled in red pencil on the answer sheet to indicate at what part of the test the man is to begin.

e. A watch with a second hand or a stop watch.

f. A whistle. The blowing of the whistle at 50-second intervals is the "change stations" signal.

g. Timekeeper's card with an arrangement of numbers as shown below.

TIMEKEEPER'S CARD				
60	60	60	60	60
50	50	50	50	50
40	40	40	40	40
30	30	30	30	30
20	20	20	20	20
10	10	10	10	10

This card will help the supervisor keep account of the 50-second intervals.

V. **Assistance needed.** Two proctors. Test supervisor will act as timekeeper.

VALVE IDENTIFICATION TEST

DIRECTIONS SHEET

I. Directions for administration of the test.

1. Before the test begins make sure that all preliminary arrangements have been attended to by checking to see that each item card matches the valve part, that stations are properly numbered, that the answer sheet at each station is properly marked with a red circle around the item number that corresponds to the station, and that a scoring key has been prepared.
2. Place the trainees around the tables so that *one and only one* man is at a station and that there are no vacant stations between the men. In case the number of men being tested is less than the number of items in the test, remove the extra answer sheets.
3. See that each man has a pencil. Have a few extra in reserve in case trainees break their pencil points.
4. Read "DIRECTIONS TO STUDENTS" aloud. Answer questions. Then say "READY, BEGIN."
5. At the end of each 50-second interval blow the whistle as the "change station" signal. Use the timekeeper's card to indicate the elapse of 50 seconds by crossing out the appropriate figure in the vertical columns each time the whistle is blown. Give the directions for starting when the second hand of your watch is on 60. Cross out the "60" at the top of the first column. The figure 50 which appears below tells you to blow the whistle when the second hand of the watch reaches 50. Continue in this manner until the test is completed.
6. During the progress of the test, make sure that the proctors continue to check the work of the students, seeing that the answers are recorded in the proper places on the answer sheet.
7. When the test is completed, collect the answer sheets and send the students to their next assignment.

II. Directions to students (To be read aloud by the test supervisor).

1. "This is a test designed to measure your ability to *identify parts of various types of valves and the function or use of each part.*
2. "Write your name, section number, and the date in the spaces on the answer sheet.
3. "Notice the large number at the top of the card in front of you. This tells you the number of the station at which you are now standing. Find this same number on your answer sheet. It has been circled with a red pencil. This circled number shows you where to start marking your answers.
4. "Above the card you will find a part from a valve. You may pick up the part and examine it if you wish. On the left side of the card are five names of parts; one of these is the name of the part before you. Select the correct name of the part and notice the number 1, 2, 3, 4, or 5, in front of it. This is the number you are to write in the first blank space beside the red circle on your answer sheet. If you are not sure of the correct name, make the best guess you can. On the right hand side of the card are five statements of part functions. One of these statements describes a function or a purpose of the valve part. Select the correct function and notice the number in front of it. This is the number you are to write in the second blank space to the right of the item number. If you are not sure of the answer, make the best guess you can. If you wish to change an answer, erase and write in your new answer.
5. "As you change stations leave the card and part where you found them. When the whistle blows, take your answer sheet with you and move TWO stations to your right, to the station with the next higher number. After you reach the highest number (29 for this test), your next station will be 1.
6. "After you finish at each station, put your answer sheet face down on the table and stand by that station until the whistle blows.
7. "Are there any questions?" (Allow time to answer any legitimate questions.)
8. "READY, BEGIN." (Give this signal when the second hand of your watch is on "60".)

III. Directions for scoring.

Scoring of the identification test may be accomplished in any one of three ways:

1. The usual method of scoring is to count the total number of correct responses given by the trainee. The number of right answers for the identification of "Name" is added to the number of correct answers given for the "Function" and this total is the score on the test. It is obvious that the total possible score is a number which is twice the number of items on the test.
2. A second method is to score each item on an "all or none" basis, that is a trainee must answer correctly both the "Name" and the "Function" on each item in order to receive credit for the item. Any item for which the right name but the wrong function is given or vice versa is counted as incorrect. The total possible score is the same as the number of items on the test.
3. The third method of scoring is a combination of the previous two. In this case one point is credited for each "Name" correctly identified, one point for each "Function" correctly indicated, and an additional point for each item that is correctly answered for both name and function. A perfect score which is three times the number of items in the test is possible under this method of scoring.

PART PLACED HERE

The valve part for this item is the piston of the pressure regulating valve (Leslie CP type).

17

Part-Name	Function
1. Main valve	1. Aligns and supports lower crosshead.
2. Spring seat	2. Opens auxiliary valve when pushed by discharge pressure.
3. Diaphragm	3. Opens main valve when forced down by steam pressure admitted through controlling valve.
4. Piston	4. Provides a means of opening and closing steam ports.
5. Slide	5. Acts as a buffer and guide for adjusting spring.

Figure 5.—Sample card and valve part used in identification tests.

CHAPTER III

WRITTEN TESTS

3A. PLANNING THE TEST

The written test involves the same necessity for planning as the performance test. Both types call for a blueprint stage prior to actual construction. The essential nature of this planning is best revealed by a statement of the questions which must be answered before beginning to write the test.

Question 1: *What knowledge is to be developed by this course of study?* As in the performance test this question is of prime importance. Every examination question prepared must be judged in terms of whether it will test that essential knowledge the men must have to do the job for which they are being trained. A good question to ask in connection with every test item is, "What will this man be expected to know and to do when he steps aboard his ship?" In order to answer such a question, you must first decide what the job is; then how it is to be done.

Question 2: *What kinds of knowledge should be tested to determine the ability of the men to do the job for which the training has been given?* Written tests should be used to do more than merely measure the ability to parrot back the information that has been taught. The question to be answered is "How well can this man apply the learned information in new, unfamiliar problem situations?" A test item which requires a man to size up a situation and decide how to solve the problem presented by the facts gives a better evaluation of a man's training than an item which merely tests a man's ability to recall the name of a piece of equipment. "Can the man reason? Can he generalize? Can he interpret information given?" These are a few of the questions which a good test should answer.

Question 3: *Having selected the areas in which knowledge is to be developed, what emphasis should the test give to each of these areas?* Decision on this question is necessary in order to determine the number and difficulty of the questions to be used in covering the various areas to be tested. In the main, your own judgment will determine the matter of emphasis. The proportion of time spent on a subject may serve as a secondary guide.

3B. TYPES OF QUESTIONS

The five most common types of questions used in written examinations are:

1. Multiple choice items
2. Matching items
3. Completion items
4. True-False items
5. Essay questions

In order to illustrate some significant test principles as well as some of the difficulties which may be encountered in preparing these various types of questions, a number of examples of good and poor practices are presented in the paragraphs which follow. The examples have been chosen from many fields. Even though the subject matter may be removed from your own particular specialty, the general principles involved will still prove applicable.

3C. WRITING THE MULTIPLE CHOICE ITEM

The examples which follow demonstrate the nature of multiple choice items and illustrate the principal cautions to be observed in connection with their preparation. Note that the first few illustrations below provide spaces for the student to mark his answer directly on his test paper alongside the question. If a separate answer sheet is used, as shown in Chapter V, Section 5B, these answer spaces next to the question need not be provided.

1. THE MULTIPLE CHOICE ITEM USUALLY CONSISTS OF A STATEMENT OR QUESTION FOLLOWED BY A SERIES OF CHOICES OR ANSWERS, ONLY ONE OF WHICH IS CORRECT.

Example—Statement to be completed:
A series motor is best suited for use on
1. lathes
2. drill presses
3. hoists
4. belt drives
5. generators ....._____

Example—Question to be answered:
Which stroke directly follows the compression stroke in a four-stroke cycle diesel engine?
1. Power
2. Intake
3. Injection
4. Exhaust
5. Scavenging ....._____

2. THERE SHOULD BE ONLY ONE CORRECT AND UNDISPUTED ANSWER.

No two ways about this:
The cathode is heated by the
1. anode
2. screen grid
3. filament
4. plate
5. suppressor ....._____

Plenty of arguments here:
The most effective plane in the Naval service is the
1. Avenger—TBF
2. Dauntless—SBD
3. Helldiver—SB2C
4. Corsair—F4U
5. Hellcat—F6F ....._____

3. EACH ITEM SHOULD TEST ONLY ONE IDEA.

Here a single, clear-cut idea is tested. Failure on this item indicates that the student has not mastered the problem of deflection in gunnery:

Deflection is always to the left when

1. it is affected only by drift
2. the wind is blowing across the the line of sight
3. the target is stationary
4. the target alone is in motion
5. the target ship and own ship are moving in the same direction ....._____

Several ideas invite confusion. If the student answered incorrectly here, it is impossible to tell if he failed to understand that part of the problem which deals with range or the part which deals with deflection:

The range decreases and the deflection is to the left when

1. the distance from the firing position to the target is reduced and no factors except drift affect deflection
2. the distance from the firing position to the target increases and the target alone is in motion
3. the distance from the firing position to the target is reduced and the wind is blowing
4. the distance to the target remains constant and the target is stationary
5. the distance to the target increases but the target and own ship are moving in the same direction

_____

4. THE CHOICES WHICH REPRESENT POSSIBLE ANSWERS SHOULD DEAL WITH SIMILAR IDEAS OR DATA RATHER THAN A VARIETY OF UNRELATED POSSIBILITIES.

These answers follow a related pattern:

An organization of two or more divisions of vessels is known as a

1. fleet
2. force
3. squadron
4. flotilla
5. division ....._____

These answers cover many relatively unrelated fields:

A ship of the light cruiser class is always

1. commanded by an admiral *(command)*
2. over 30,000 tons *(weight)*
3. used as part of all task forces *(tactics)*
4. armed with 16-inch guns *(armament)*
5. designated CL *(symbols)*....._____

5. MULTIPLE CHOICE QUESTIONS MAY BE USED EFFECTIVELY TO PRESENT PROBLEMS INVOLVING REASONING BASED ON KNOWLEDGE. SUCH ITEMS REQUIRE THE STUDENT TO USE HIS KNOWLEDGE RATHER THAN TO DEMONSTRATE HIS MEMORY FOR SMALL, DETAILED FACTS.

This question calls for reasoning based on an understanding of structure and function:

If a 40 MM stops firing with a round on the tray and the rammer shoe in the forward position, there is most probably a broken

1. firing pin
2. tray catch lever
3. star wheel catch arm
4. trigger catch lever
5. star wheel catch spring ____

This question tests only the student's memory for names:

The rotating core of an electric motor is known as the

1. field
2. centroid
3. commutator
4. armature
5. brush .....____

6. AVOID QUESTIONS WHICH CAN BE ANSWERED SOLELY ON THE BASIS OF INTELLIGENCE, OR THE MOST GENERAL KNOWLEDGE, WITHOUT NEED FOR HAVING ANY SPECIFIC KNOWLEDGE OF THE SUBJECT TESTED.

To answer here you must know your subject:

Arcing at the breaker points is prevented by the

1. condenser
2. coil
3. distributor
4. shielding harness
5. potentiometer .....____

Any sensible person can narrow these choices down to 1 or 5 without knowing a thing about physics. Very general knowledge points to 5 as the answer.

In the ocean, the greatest water pressure will be found at a depth of

1. one mile
2. two miles
3. three miles
4. four miles
5. five miles .....____

7. AVOID ASKING QUESTIONS ON TRIVIAL DETAILS AND USELESS SUBJECTS.

Useful knowledge:

Hydraulic systems using mineral base fluids are flushed and cleaned with

1. soap solution
2. kerosene
3. carbon tetrachloride
4. alcohol
5. caustic soda solution _____

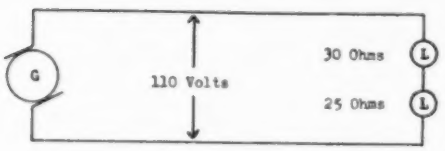
Trivial. Not much practical value in knowing this:

The term which best describes the science of using ordnance material is

1. ballistics
2. cannonade
3. boresighting
4. gunnery
5. tactics .....____

8. INGENUITY IN THE CONSTRUCTION OF ITEMS ADDS TO THEIR INTEREST AND USEFULNESS. DIAGRAMS, GRAPHS, AND PICTURES ADD REALITY TO TEST ITEMS AND BRING THEM CLOSER TO PRACTICAL SITUATIONS. QUESTIONS BASED ON PHOTOGRAPHS OF EQUIPMENT OR WORK BEING DONE BRING THE TEST CLOSER TO AN ACTUAL PERFORMANCE SITUATION.

a. *A realistic problem is presented below through use of a diagram:*

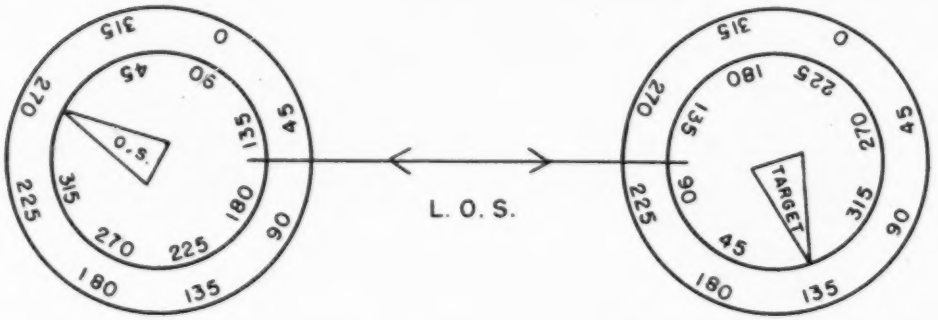


The current through the combination in the diagram at the left is

1. 0.2 amperes
2. 0.5 amperes
3. 2.0 amperes
4. 5.0 amperes
5. 20.0 amperes

b. *Four realistic problems are presented in the next question through the use of a diagram. Some really significant knowledge may be tested in this way.*

In the diagram below, the approximate values of Cr, Co, A, and Br are



Ct

1. 000°
2. 005°
3. 135°
4. 175°
5. 275°

Co

1. 000°
2. 005°
3. 135°
4. 175°
5. 275°

A

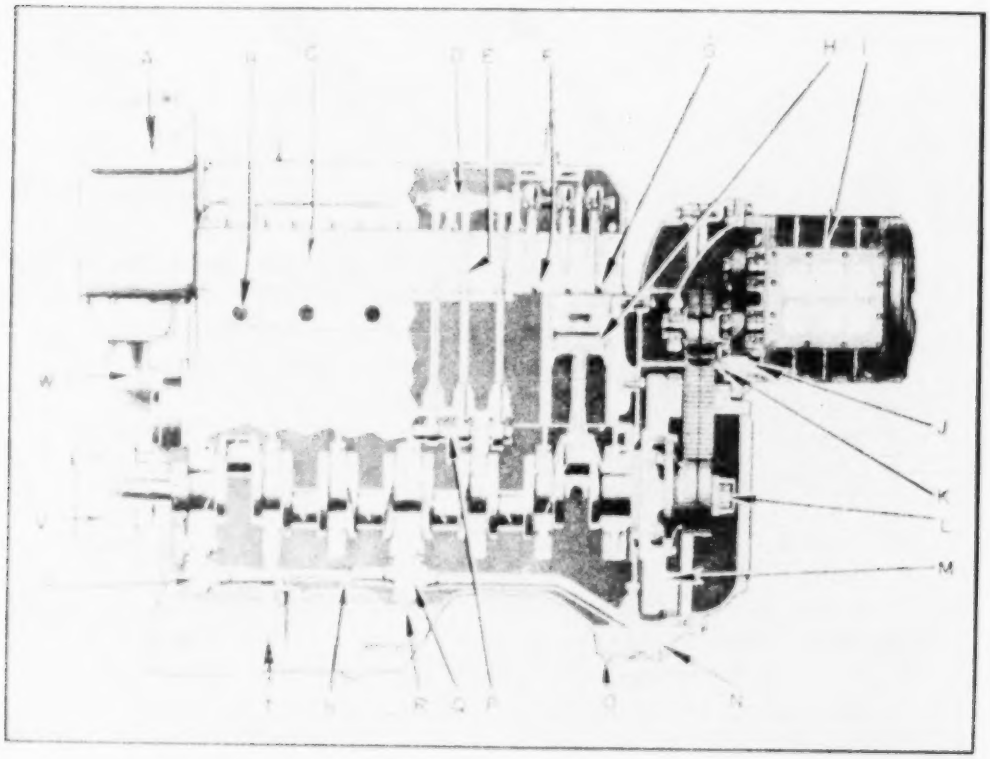
1. 060°
2. 100°
3. 150°
4. 215°
5. 235°

Br

1. 060°
2. 100°
3. 150°
4. 215°
5. 235°

c. *Many practical questions may be based on a photograph of equipment if reproducing facilities are available. A long series of questions has been developed in connection with the diesel engine pictured below. Only a few are presented to illustrate the procedure.*

Refer to this picture in answering items 124 through 139.



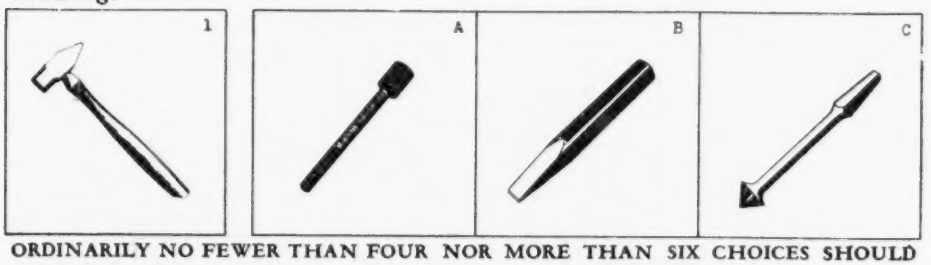
124. The part of the engine shown at P is the
1. balance shaft
2. crankshaft
3. camshaft
4. vibration damper
5. rocker arm shaft

125. The line at W carries
1. sea water
2. fresh water
3. lubricating oil
4. air
5. fuel oil

126. An intake valve is located at
1. B
2. C
3. E
4. F
5. G

127. If this engine is used for propulsion there will be attached at L a
1. clutch driving drum
2. clutch plate
3. main drive shaft
4. flexible coupling
5. vibration damper

d. *It is entirely possible, by using a series of drawings or photographs, to develop a multiple choice item which eliminates all words in the examination except for the directions. In the item below, the problem is to decide which one of these drawings labeled A, B, and C is most closely related to the first drawing.*



9. ORDINARILY NO FEWER THAN FOUR NOR MORE THAN SIX CHOICES SHOULD BE USED. FIVE CHOICES WILL GIVE GOOD RESULTS.

With five choices, the answer is hard to guess. There is only one chance in five for a correct guess:

Inside threads on a pipe are cut with a

1. tap
2. chuck
3. die
4. reamer
5. knife edge

With two choices the answer is easier to guess. There's a chance of guessing correctly on half the questions:

A diesel engine operates on

1. gasoline
2. fuel oil

10. AS MUCH OF THE ITEM AS POSSIBLE SHOULD BE INCLUDED IN THE INTRODUCTORY STATEMENT.

Short, clear, easy to read:

If the primary of a transformer has a current flow of 40 amperes and a voltage of 220, and the voltage in the secondary is 550, the secondary current is

1. 15 amperes
2. 16 amperes
3. 19 amperes
4. 30 amperes
5. 40 amperes

Long, repetitious, confusing:

Sixteen amperes

1. will flow in the secondary current if 40 amperes flow in the primary of a transformer with a voltage of 220, and if the voltage in the secondary winding is 550.
2. will be the resulting current in the secondary if 30 amperes flow in the primary of a transformer with a voltage of 220, and if the voltage in the secondary winding is 550.
3. will flow in the secondary current if 40 amperes flow in the primary of a transformer with a voltage of 110, and if the voltage in the secondary winding is 550.
4. and 5. (More of the same long winded stuff.)

11. KEEP THE LANGUAGE AND PUNCTUATION CLEAR.

Clear:

When it is 1600 on 21 March in San Francisco (Zone 8), the date and time in Manila (Zone minus 8) is

1. 2400 March 20
2. 0800 March 21
3. 2400 March 21
4. 0800 March 22
5. 1600 March 22

Cloudy:

In Manila (Zone minus 8) when it is 1600 March 21st in San Francisco (Zone 8) it is

1. 2400 March 21
2. 0800 March 22
3. 1600 March 22
4. 0800 March 22
5. 2400 March 20

12. THE CHOICES WHICH FOLLOW THE INCOMPLETE INTRODUCTORY STATEMENT SHOULD COMPLETE THAT STATEMENT IN A GRAMMATICAL FASHION.

Consistent and grammatical:

When heated, most lubricants become

1. less corrosive
2. adhesive
3. more viscous
4. thinner
5. more compressible

Several choices here have no grammatical relation to the introductory statement:

The step in a hydroplane float is intended to

1. add buoyancy
2. it reduces landing speed
3. it breaks suction
4. has no effect
5. streamlining to cut drag

13. BE CAREFUL WHEN USING "A" OR "AN" AS THE FINAL WORD IN THE INTRODUCTORY STATEMENT. IT MAY GIVE A CLUE TO THE ANSWER.

No clue here:

The instrument used to measure resistance in a circuit directly is

1. a voltmeter
2. an ohmmeter
3. a watt meter
4. a pitometer
5. a potentiometer

Only the right answer, "an ohmmeter," makes grammatical sense here:

The instrument used to measure resistance in a circuit directly is an

1. voltmeter
2. ohmmeter
3. wattmeter
4. pitometer
5. potentiometer

14. IF ANSWERS ARE TO BE WRITTEN ON THE TEST PAPER ITSELF, THEN ANSWER SPACES SHOULD BE BROUGHT OUT TO THE RIGHT OF THE QUESTION AND ARRANGED IN A COLUMN ON THE TEST PAPER FOR QUICK AND SIMPLE SCORING.

Answer space separated from question:

A rheostat is a
1. transformer
2. generator
3. variable resistor
4. motor
5. spark gap ....._____

Answer space runs into question:

The .50 caliber Browning machine gun is operated by_____
1. gas
2. compressed air
3. the recoil
4. hydraulic action
5. blow back

15. IF CHOICES ARE LISTED IN A COLUMN, THEY WILL PROVE EASIER TO READ AND LESS CONFUSING TO ANSWER.

Choices arranged in a column are easily read:

If 1000 feet of #14 R.C. wire cost $19.75, then 825 feet will cost
1. $15.29
2. $16.29
3. $16.92
4. $23.49
5. $23.94

Choices which are run together may cause confusion and careless errors:

A one-degree change in blade angle affects engine rpm by (1) 30-50 rpm, (2) 70-100 rpm, (3) 110-130 rpm, (4) 140-160 rpm, (5) 170-200 rpm.

16. SOME ADDITIONAL POINTS ON THE MULTIPLE CHOICE ITEM.
   a. This type of item doesn't attempt to test the student's ability to organize and present his knowledge in his own language, which may sometimes represent an important factor in success on the job.
   b. The multiple choice item tests the student's *recognition* of the correct answer from among a series of possible choices. It doesn't indicate with certainty whether the student would have *recalled* the correct answer without any hints.
   c. Use plausible choices. The trainee should be required to pick the *best* answer from among several that are nearly as good. Where the question requires computation use the kind of errors that occur most frequently in students' work.
   d. Scatter the position of the correct answers. Choices 1, 2, 3, 4, and 5, should appear as correct answers with almost equal frequency. Avoid any pattern of placing the correct choices.
   e. In each item where the answers are numerical values it is a good practice to arrange the choices in order of magnitude. Note example 6, 8a, 8b, 8c-126, 10, 11, and 15.

3D. MATCHING ITEMS

Matching questions generally include two lists of related words, phrases or symbols. The student is required to match *each* item in one list with *some one* item with which it is most closely related in the second list.

There are many effective uses to which the matching item may be put. The following illustrations, taken from several fields, show a variety of means for getting at significant knowledge through use of this type of test item.

1. MATCHING ITEMS ARE USED IN TESTING KNOWLEDGE OF THE FUNCTIONS OF PARTS, MACHINES, TOOLS, MATERIALS, ETC.

Indicate the part of the gasoline engine which carries the functions listed under column B. To the right of each function in list B write the number of the one most closely related part in list A.

A	B
1. Cam	(44) Collects exhaust gases from exhaust valves or ports.... ____
2. Carburetor	(45) Admits fuel mixture directly to the cylinder..... ____
3. Crankcase	(46) Regulates the opening and closing of the intake valve.. ____
4. Crankshaft	(47) Ignites air and fuel mixture____
5. Cylinder	(48) Effects a gas tight seal between piston and cylinder.. ____
6. Exhaust manifold	(49) Provides a reservoir for lubricating oil ..... ____
7. Flywheei	(50) Translates reciprocating motion of the pistons to rotary motion ..... ____
8. Intake manifold	(51) Distributes air and fuel mixture to intake valves... ____
9. Intake valve	
10. Piston ring	
11. Spark plug	
12. Water pump	

2. MATCHING ITEMS MAY BE USED IN THE CLASSIFICATION OF PARTS, MACHINES, PERSONNEL, ETC.

Note the directions in the following example which indicate that any answer may be used more than once. Note also that this is one of the exceptional situations in which the number of items in the suggested answer list may be less than the number of items in the question list.

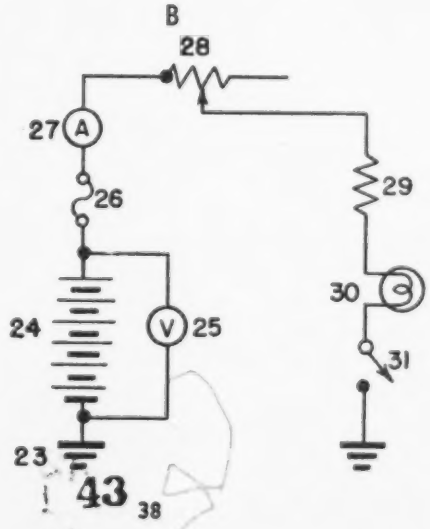
In list A are found the names of various types of airplane wings. To the right of each airplane named in list B write the number of the one most closely related item from list A. Any answer contained in list A may be used more than once.

A	B
1. Gull	(25) PBY "Catalina" .....____
2. High	(26) SBD "Dauntless" .....____
3. Low	(27) SO3C "Seagull" .....____
4. Mid	(28) C-47 "Skytrain" .....____
5. Parasol	(29) "Sunderland" .....____
	(30) F4U "Corsair" .....____
	(31) OS2U "Kingfisher" .....____
	(32) PBM "Mariner" .....____

3. MATCHING ITEMS MAY BE USED TO SHOW RECOGNITION OF TECHNICAL SYMBOLS.

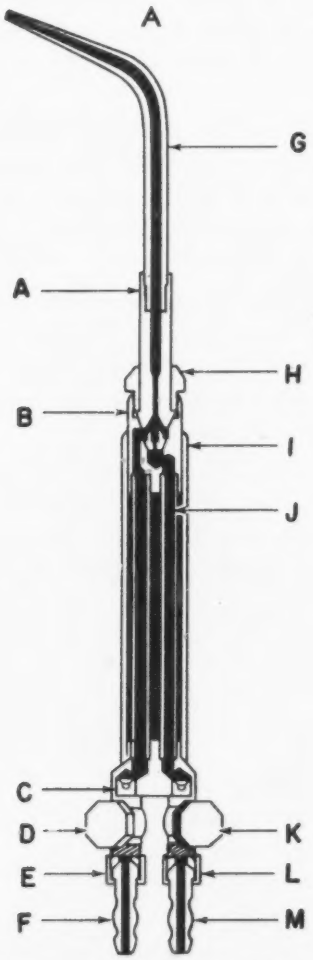
Identify the electrical symbols illustrated by the circuit drawing under list B. In the space following each number in the list to the right of column B, print the number of the one most closely related item in list A. The numbers refer to the circuit shown.

A	B
1. Ammeter	23. ____
2. Battery	24. ____
3. Fixed resistor	25. ____
4. Fuse	26. ____
5. Generator	27. ____
6. Ground	28. ____
7. Lamp	29. ____
8. Relay	30. ____
9. Rheostat	31. ____
10. Switch	
11. Voltmeter	
12. Wattmeter	



4. MATCHING ITEMS MAY BE USED IN THE IDENTIFICATION OF PARTS OF MACHINES, TOOLS, ETC.

Certain parts of the welding torch, pictured below, have been lettered under A. Place the letter of each part in the space to the right of the appropriate name in list B.



B

(9) Oxygen hose connection nut .....____

(10) Acetylene hose connection nut .....____

(11) Oxygen needle valve .....____

(12) Acetylene needle valve .....____

(13) Mixing head .....____

(14) Oxygen tube .....____

(15) Acetylene hose gland .....____

(16) Oxygen hose gland .....____

(17) Tip .....____

(18) Torch head .....____

5. IN MUCH THE SAME WAY MATCHING ITEMS CAN BE USED TO TEST KNOWLEDGE AS TO:

a. Definition of terms.
b. Cause and effect.
c. Units to which various parts belong.
d. Duties of various personnel, organizational groups, etc.
e. Solution of mechanical, mathematical, and practical problems.
f. Composition or structure of mechanical parts, materials, etc.
g. Procedures used in various operations.

6. SOME BASIC RULES FOR CONSTRUCTING MATCHING ITEMS.
   a. Keep directions clear and the form simple enough to make for ease in answering and scoring.
   b. A matching question should generally deal with only one subject.
   c. As a corollary to rule b., above, don't mix numbers, names, and dates in a general list of suggested answers.
   d. Don't give clues, such as statements in the column to be matched which end in "a" or "an" when only a limited number of items in the answer column can fit such adjectives. Mixing plural and singular nouns in the answer column may also give a clue.
   e. Don't have items and answers in the two lists in about the same order. Mix up the questions and answers so that they appear in a random order. (It's usually a good idea to list the answers in alphabetical order, or in order of magnitude. When a picture or diagram is used the key numbers or letters should be arranged in order.)
   f. At least five but no more than twelve questions to be answered should be included in each matching item.
   g. The column containing suggested answers should contain three to five more elements in it than the column with items to be matched (unless the same answers may be used for two or more items).
   h. If any of the suggested answers listed may be used to answer more than one question, the directions should clearly state that "any answer may be used more than once." (For illustration see paragraph 2 above.)
   i. All parts of the matching unit should appear on the same page.

7. EXAMPLE: WHAT NOT TO DO.

Match the following items in each list.

A	B
1. Prosign for "repeat."	1. J
2. Prosign for "verify and repeat."	2. $\overline{\text{IMI}}$
3. "Mayday" is used as a	3. distress signal
4. Group count of a message.	4. operating signals
	5. GR 165
	6. 121445

The foregoing illustration violates many of the basic rules presented in paragraph 6 above.

In the above example are you required to match each item in list A with some item in list B, or vice versa? The directions are not specific on this point and the student might incorrectly attempt to match each item in list B with some item in list A. This possibility of confusion is increased by failure to provide specific lines on which to write the answers. The student would probably place answers haphazardly in the center of the page alongside the items in list A. Such a careless arrangement would increase the difficulty of scoring and invite scoring errors. By

reversing columns A and B and placing answer lines in a straight column at the right margin of the page, both answering and scoring will be simplified.

The example given deals with three separate subjects in radio communications; namely, prosigns, distress signals, and word counts. This makes it easy to figure out which suggested answers under list B must relate to particular questions under list A. The group count question undoubtedly deals with numbers. Therefore, *only* 5 and 6 under list B can apply to question 4 under A with reference to group counts. Prosigns undoubtedly refer to the code signs J and $\overline{\text{IMI}}$. This leaves only "distress signal" and "operating signals" as possible answers for the "Mayday" question. Based on this altogether accurate reasoning, each question has but two possible answers and the chance of correctly guessing the answer is considerably increased.

The group count question in the above example shows how poorly numbers and general information answers mix. Questions which call for number or personal name answers are easily separated from more general information questions. It thus becomes easier to reason through to an answer without actually having the required knowledge.

In the example used, the answer to item 3 in list A is pretty obvious. This item reads, " 'Mayday' is used as a," and the possible answers from list B have already been narrowed down to "distress signal" and "operating signals." Which of the two is the correct answer? It must certainly be "distress signal." It makes sense to say, " 'Mayday' is used as a distress signal." It makes no grammatical sense to say, " 'Mayday' is used as *a* operating signals." The "a" doesn't fit; nor does the plural "signals."

In the above example, the first two questions deal with the first two answers; the next question is related to the next two answers, and the last question is related to the last two answers. This procedure gives additional clues to the proper answers. One way to avoid this difficulty is to place the suggested answers in alphabetical or numerical order and then to place the "question" column in a random order.

In the above example, only four questions are asked resulting in a poor item. It is possible to compress quite a few questions in a small space when the matching type test is used. Unlike the multiple choice item in which each question has its own separate series of possible answers, all the suggested answers in the matching type item serve as possibilities for every question. It would, therefore, be uneconomical to have fewer than five questions. Having more than twelve questions, however, would probably make the item too long and confusing. The additional possible answers serve only to increase the difficulty of the problem. They should, therefore, be sensible enough to "cross up" the person who is just guessing. In the example above, not enough of these possible, yet incorrect, answers are given to reduce the element of guessing and help weed out the fellows who are "shooting blind."

3E. COMPLETION ITEMS

This type of test item generally consists of a statement with one or more words omitted. The missing word or words must be supplied from memory. Such an examination item is particularly suitable for testing memory for material which must be recalled in a precise way, such as technical terms that must be known, abbreviations, weights, measures, tolerances, and the like. It is also suitable for the presentation of mathematical problems. However, there are many dangers to be avoided.

1. ALWAYS BE SURE THERE IS ONE, AND ONLY ONE, ANSWER.

There is only one possible word which can be inserted in the example below. It is "torque."

The wrench used to measure the amount of twisting force being applied to a nut is called a _____ wrench.

Many answers, from "magazine" to "oil tank," would fit this item. It's much easier than you think to get fouled up by some surprise answer that is perfectly correct.

One should never smoke in a ship's _____.

2. IT IS GENERALLY BEST TO USE ONLY ONE BLANK IN A SINGLE SENTENCE.

This is straightforward:

A generator is a device for converting mechanical energy into _____ energy.

For mind-readers only:

A _____ is a device for converting _____ energy into _____ energy.

3. TRY TO KEEP THE BLANKS NEAR THE END RATHER THAN THE BEGINNING OF THE SENTENCE.

Idea to be tested stated first.

A device used to prevent an overload in an electric light circuit is a _____.

A second reading is necessary to understand what is wanted.

A _____ is used to prevent an overload in an electric light circuit.

4. OMIT ONLY THOSE KEY WORDS WHICH THE STUDENT SHOULD KNOW. DON'T ASK FOR RECALL OF SOME TRIVIAL DETAIL.

Important:

In case of man overboard, break out and lower the numeral ____ flag.

A poor omission:

In case of man overboard, break out and lower the _____ FIVE flag.

5. ARRANGE THE FORM OF THE QUESTION SO THAT ANSWERS MAY APPEAR IN A COLUMN AT THE RIGHT. THIS WILL INCREASE SCORING SPEED AND ACCURACY.

An aerial torpedo is made up of ____ main sections .....

6. IT WILL OFTEN BE POSSIBLE TO USE A STRAIGHTFORWARD QUESTION IN PLACE OF AN INCOMPLETE STATEMENT. THIS ARRANGEMENT MAY OFTEN PROVE JUST AS EFFECTIVE AND LESS CONFUSING.
   How many main sections are there in an aerial torpedo? .....——

7. SOME ADDITIONAL RULES:
   a. Don't omit the verbs in the sentences.
   b. Do not copy statements directly from textbooks to make a completion item.
   c. Each blank should call for a one-word response rather than a group of words. If you must omit a phrase or other group of words, indicate it by using a separate line for each omitted word.
   d. Avoid being too brief. Make the statement complete enough so that there is no doubt as to its meaning.

3F. TRUE-FALSE ITEMS

This type of test item consists of a single statement which is to be marked true or false. The pitfalls and shortcomings of this type of item warrant careful consideration before it is used very extensively. Since there are only two alternative answers for this type question, it encourages guessing. Half of the questions might be answered correctly without any knowledge of the subject. In addition, it is very difficult to make a statement which is absolutely true or absolutely false without giving some hint as to the correct answer. Finally, a relatively large number of such questions must be used in an examination if there is any expectation that the test will separate the good students from the weaker students. Some rules for the construction of such items are illustrated below.

1. DON'T HAVE ONE PART OF THE ITEM CONTAIN A TRUE IDEA AND ANOTHER PART A FALSE IDEA. MAKE IT ALL TRUE OR ALL FALSE.

All true:	*Part true, part false, all confusing*:
A Navy admiral is considered of flag rank.	A Navy admiral wearing four stripes on his uniform is considered of flag rank.

2. AVOID DOUBLE NEGATIVES OR INVOLVED STATEMENTS.

Properly stated:	*What does this mean*:
Before working on any electrical or radio equipment one should be certain that all circuits are de-energized.	One should not work on any electrical or radio equipment if he is not sure that no circuits are energized.

3. AVOID USING THE WORDS "ALL," "ONLY," "NEVER," "ALWAYS," "GENERALLY," OR "USUALLY" IN THE STATEMENT UNLESS THESE WORDS ARE AN IMPORTANT PART OF THE STUDENTS' UNDERSTANDING. THEY GIVE A HINT AS TO THE ANSWER.

Statements with the words "all," "only," "never," "always" are usually false—as are the following:

All airplane propellers are made of aluminum.

Torpedoes are always dropped within 1000 yards of the target.

Never run an airplane engine in a hangar.

Only casein glue may be used in joining wooden airplane parts.

Statements with the words "generally" or "usually" in them have a tendency to be true—as are the following:

Sea plane floats are generally equipped with rudders.

Small tears in airplane fabric can usually be repaired by patching.

In general, destroyers are faster than battleships.

4. SOME ADDITIONAL RULES.
   a. Make about half the items true and half of them false.
   b. Have the true and false items thoroughly mixed, but not in any set pattern.
   c. Avoid tricky language or trivial technicalities.
   d. Do not make true statements consistently longer or shorter than false statements.

3G. ESSAY QUESTIONS

The type of question which calls on the student to describe, compare, discuss, or explain some aspect of the subject he is studying has a very limited use in Navy enlisted schools. It might be useful when the student's ability to organize facts and ideas and then to reason from them is of prime importance. Otherwise, the shortcomings inherent in this type of examination question make it advisable to use the kinds of test items discussed in the preceding paragraphs. *The first of these shortcomings* is evident from the *relatively long time it takes to write an essay answer,* thus limiting the number of points which can be covered in a reasonably short examination. *Equally serious is the difficulty of scoring* the essay answer in a fair and accurate way. *Finally there is a tendency for poorly framed essay questions to invite bluffing* of the sort in which literary skill counts for more than real knowledge. Attention to the rules which follow will reduce rather than eliminate these weaknesses.

1. THE QUESTION SHOULD INDICATE THE GENERAL OUTLINE FOR THE ANSWER. SETTING OUT THE SPECIFIC POINTS TO BE DEALT WITH WILL REDUCE THE OPPORTUNITY FOR BLUFFING. This, in effect, changes the general essay question to a short answer essay question. It becomes much easier for both students and instructors to handle.

This question requires a clean-cut answer based on real knowledge. The points to be scored are very specific:

Explain the steps in the preflight inspection of the Pratt and Whitney Twin Wasp R-1830 airplane engine, giving
1. parts to be inspected.
2. what difficulty would be looked for in each part.

These questions are a cinch for a wind-bag. Besides, no two answers will be even remotely similar and scoring will be a headache:

Explain the operation of a Pratt and Whitney Twin Wasp R-1830 airplane engine.

Discuss the Navy rating system.

Discuss the supervisory duties of a radioman.

2. AVOID WRITING QUESTIONS IN ELABORATE AND COMPLICATED LANGUAGE, AS WELL AS QUESTIONS IN WHICH THE QUALITY OF THE ANSWER WILL DEPEND ALMOST ENTIRELY ON THE STUDENT'S MASTERY OF LANGUAGE.

No five dollar words given or asked for:

State briefly the reasons a propeller blade which is in motion
1. tends to turn toward high pitch.
2. bends forward at the tips.
3. bends in the direction opposite to the direction of rotation.

A mighty good propeller man could bilge on this one:

Discuss three major dynamic forces which become operative when the propeller is energized by revolutions of the crankshaft.

3. SOME GENERAL RULES FOR SCORING.
   a. Score all the papers on one question at a time. Grade the answer to the question—not the man. Compare the quality of the answers given by the various men to a given item; this cannot be done by grading more than one question at a time.
   b. Set your standards. Before reading any student papers write out a model answer. Check the essential and important points which should be included in the discussion. Determine the values to be assigned to each point.
   c. Read over a number of student answers before scoring the papers to verify your model and the point values assigned.
   d. Check each student answer against the verified model.

3H. HOW LONG SHOULD THE TEST BE?

As a practical matter, the length of a test is often determined by the amount of time available for testing. A daily quiz may be limited to four or five items while a final achievement examination may contain two hundred questions and last for two hours or more.

Despite this necessary variability, it is of the utmost importance that the instructor realize the significance of tests of varying length. A test of ten or fifteen true-false or multiple choice questions cannot be relied on to give a true picture of the achievement of individual members of the class. A very much longer test is needed before the instructor can say with confidence that he has measured the relative accomplishment of his students.

Exactly how long the test must be for reliable results cannot be stated in advance. As a general rule, from seventy-five to one hundred questions would be a minimum for a test designed to estimate the relative achievement of students in the average subject. This refers to questions of the multiple choice, matching, or completion form. If true-false questions are used a much greater number would be required, perhaps one hundred and fifty questions.

This is not to deny the usefulness of short daily or weekly tests. The only caution to observe is to make no estimate of the achievement of a student on the basis of a single short examination. However, the results of a number of such tests might well be considered as if they constituted a single, long test. By considering a whole series of such short examinations, a fairly reliable measure of student achievement would be obtained. However, a short test shows what the man can answer on the day he takes it. Adding the results of short tests will tell you how much a man knew at some time during the training period. The sum total will not tell you how much he retains at the end of the course.

3 I. CHECK LIST FOR BUILDING THE WRITTEN ACHIEVEMENT TEST

Discussion of the written test to this point has been confined to two main subjects: (1) preliminary planning of the test; (2) writing various types of items.

There are other important procedures and rules incidental to the construction of written achievement tests which must be noted.

1. *In a single test, use no more than two or three different types of items.*
2. *Prepare more items than will actually be used in the final form of the test.* This will leave a reserve so that weak or faulty items may be eliminated.
3. *Include some reasonably easy items and some very difficult items:*

Each test should include enough easy items to tell whether a man has mastered the minimal essentials of the training. It is unlikely that any student will be completely ignorant of the subject tested. Each test should contain a large number of items of moderate difficulty to distinguish between the achievement of average and good trainees. The test should include a few items of increasing difficulty that only the best students will be expected to answer. Items that everyone answers or everyone misses do not contribute to the measuring effici-

ency of the test. By including both simple and knotty problems along with those of average difficulty, a truer picture of the relative achievement of the men will be obtained.

4. *Whenever possible, arrange the items in the order of estimated difficulty, starting with the easier items and finishing with the more difficult ones.*

In tests which cover a single topic or closely related topics it is desirable to set up the items in order of their estimated difficulty so that by answering the easier items first the student has a feeling of confidence and builds up the proper mind-set for taking the test. However, a more important reason for such an arrangement of easy to difficult items is that a more accurate measure of a man's ability is obtained if he is permitted to complete that portion of the test for which he knows the answers, before he is confronted with the more difficult items which are apt to cause confusion in his thinking. The inclusion of difficult items at the beginning of the test is apt to retard a man's progress and prevent him from reaching the easy items which he could have answered had they been placed near the beginning.

Where a test covers a number of somewhat different topics (or where scores on different sections having to do with various phases of the instructional program are wanted), it is desirable to set up the items according to difficulty within the separate sections. In this case it is advisable to either (1) set up time limits for the sections, or (2) make the directions such that the trainees will go through the entire test answering first the questions they are sure of, and then return to work on the more difficult items.

5. *Organize the test items so that scoring is simple, rapid, and accurate.*

Except for essay type questions, all test items can be arranged so that answers will appear in straight columns on the test paper. This will enable you to score papers quickly and with a minimum of error by the use of a master test copy or scoring key. Of course, if a separate answer sheet is used no provision need be made for answer spaces on the test paper.

6. *Prepare clear and precise directions for giving the tests.*

Two sets of directions should be prepared. One set of directions is needed for the *examiner* to set up the mechanical process to be followed in the administration of the test. The second set is prepared for the *students* to follow in taking the test. These directions should be printed at the beginning of the test. The student should know exactly what to do after reading these directions without further oral explanation.

7. *If at all possible, try out the test before giving it in its final form.*

Giving it to a small group of advanced students, a couple of instructors, or any other group familiar with the subject will help you iron out unsuspected flaws in the framing of the questions and the directions. It will also help you in

determining the proper length of the examination. Make the test long enough to cover the major aspects of the subject on which you are examining. Make the test short enough so that a majority of the men will finish it in the time allotted. This latter caution is important since on most achievement tests speed should not be considered as a factor in success on the test. If the test is so long that only the fastest readers can finish, then they secure an advantage over slow readers. The final score in such a test will tend to represent a student's ability in reading as much as his knowledge of your specific subject. Unless reading skill is specifically important to achievement in your subject, it would be best to minimize its effect as a factor in the test score.

8. *Check with other instructors and specialists on every phase of your work in building the test.*

Have you covered and given proper emphasis to the important terms, principles, and theory with which the men must be acquainted? Does the test adequately cover the operations and procedure basic to your subject? Are the questions framed and organized properly? Will administration and scoring of the test be simple? These are some of the questions that should be raised in conference with your associates. Remember, you often overlook the errors you make yourself though they are quite evident to others.

CHAPTER IV

ADMINISTRATION OF TESTS

4A. A GENERAL RULE

It takes more than a good ruler to measure a yard. How the ruler is handled plays a big part in the accuracy of the end result. In the same way, a good test can produce good results only when used with care and skill. Stated very broadly, expert administration of any test requires that it be given to all groups (1) *in the same way and* (2) *under the best possible conditions.*

1. Giving the test in the same way.

A test is not a handicap race. Each man in the class and each separate class should have the same chance to make a good score. Burdens should not be imposed on one individual or class which are eliminated for other individuals or classes. Uniformity in the administration of a test is absolutely essential if students' scores are to be compared accurately and fairly. Yet many testers, while accepting this rule in principle, violate it in practice. In written tests, careless little distractions are allowed to break in on the thinking of students. Or the time for taking the test is varied so that an hour time-limit becomes an hour more or less. In performance tests even the test items may be changed somewhat from student to student, or the tools and equipment provided may vary in quality sufficiently to affect the end result. A good test deserves precise and uniform administration. Without these, the best tests are of no value.

2. Giving the test under the best possible conditions.

While it is of the utmost importance to give the same test in the same way at all times, such uniformity is but one of the virtues. After all a test situation may be uniformly good, and then again it may be uniformly poor. Beyond uniformity, there is still the necessity for providing the kind of test situation which will insure the most efficient work by the students taking the test. Compare the performance test procedures described below.

STUDENT A

Good Test Situation

a. Room well ventilated.
b. Testing station well lighted so that he sees clearly the mechanism he is to disassemble.
c. Enough room between this student's station and the next so there is no distraction.
d. The student has all the tools needed for the task and enough space in which to work.
e. The test assistant has explained clearly just what is to be done.
f. The student knows on what basis he is being graded.

STUDENT B

Poor Test Situation

a. Room close and stuffy.
b. No outside light; electric lights very dim.
c. Testing station is set up at one end of a shop where other students are working. Plenty of distraction.
d. Testing assistant waits until the student asks for the necessary tools before getting them from the tool cage.
e. A visitor comes in with the Executive Officer and watches the student take the test.

What is the result? Student A works calmly and quickly. All conditions are in his favor. Student B is distracted. He has trouble seeing the intricate parts of the mechanism. He wastes time waiting for the assistant. The presence of a visitor and a school officer makes him nervous. He fumbles with his tools, is undecided, and fails to complete the job on time.

The work problem was the same in each case, yet the testing situation made the task more difficult for Student B than for Student A. Not only will the instructor be unable to compare these students accurately and fairly, but he really will not know how competent Student B is. Failure to make the two situations uniform prevents an adequate comparison. Failure to have a good test situation for B prevents an adequate judgment as to his real ability.

4B. SPECIFIC POINTS FOR THE WRITTEN TEST

The foregoing example concerns the administration of a performance test. The same principles would hold true for a written test. Following is a list of some specific points to be checked in preparing for and administering written tests:

1. **Conditions in the examination room.**
   a. Schedule cleared, no other group breaking in during test time.
   b. Training aids affecting the test removed or screened from sight.
   c. Adequate working space, tables, arm-chairs, (lapboards as a last resort); seating at adequate distance.
   d. Adequate light, ventilation (and heat).
   e. Public address system, if needed for large rooms.
   f. Blackboard and chalk, if time is to be posted.

2. **Readiness of testing materials.**
   a. Adequate supply of test sheets or booklets, spares for imperfects, answer sheets.
   b. Supplementary materials—scratch paper, allowed references, spare pencils.
   c. Examiner's equipment, timing device if required, manual of directions.

3. **Preparation of proctors.**
   a. Familiarity with procedures prescribed for the test to be given.
   b. Responsibilities in seating trainees, passing out and collecting testing materials, supervising trainees while directions are being given, assignment to dennite sectors.
   c. Proctors may give assistance in showing trainees the "mechanics" of taking the test, but may not give help on the content of the test.

4. **Preliminary instruction to trainees.**
   a. Tell them at preceding class, or at muster, what materials to bring to exam, such as slide rule, drawing instruments, logarithm tables, etc.
   b. Upon their entering the room, direct them to seats according to plan.

5. **Procedure in test administration.**
   a. Have an established procedure; if the same test is given to two or more groups, follow it exactly for each one.
   b. If necessary, give preliminary directions before distributing test material.
   c. Distribute test materials according to plan.
      (1) If separate answer sheets are used, they may be inserted in test booklets in advance, or separately distributed. If separately distributed, they should be issued before the test booklets.
      (2) If special equipment, such as electrographic pencils, slide rules, log tables, etc., is to be used, it should be issued before the papers.
      (3) Check distribution to be sure no duplicates are issued and each man has the required material.
   d. Give directions for identification of papers or answer sheets. Specify the information to be given, where it is to be placed on the paper. PRINT names, etc. (Proctors check correct performance.)
   e. Read directions (previously prepared) for taking the test. These should cover:
      (1) How to work on the test: Statement of purpose; time to be allowed and any time limits which may be set for sections within the test; what to do on each type of items included in the test; whether to guess or omit an item when not sure of the answer; advice to go through the test answering known items first, then giving more time and study to the more difficult items; whether to make necessary computations or diagrams on test margin, back of answer sheet, scratch paper.

(2) How to mark the answers: If on the test paper, whether by writing letter or number in a blank, crossing out, circling; if on a separate answer sheet, the procedure required for the specific form of sheet used.

f. Give directions on what to do if the test is poorly printed, has a page missing, pencil breaks, or there is other need for special action during the examination, and if the test is completed short of the time limit.

g. Answer any questions on what to do.

h. Give starting signal and record time, or start the watch. If time is to be posted, mark it on the blackboard.

i. Have proctors check to be sure the trainees are following directions and correct the individuals who are off on the wrong foot.

j. Check the number of trainees taking the test and the number of test booklets issued to be sure there are no strays.

k. Maintain quiet supervision of the examination. Be alert for indications of trainees who need attention, and for distracting influences which should be controlled or corrected. Proctors should move quietly about within their assigned sectors, observe the work and progress of trainees, report any unusual circumstances to the examiner in charge, change location of trainees where circumstances such as being too near to a radiator or an open window, or bad lighting indicate such need. However, if the test has strict time limits, changing location of trainees should be done only when there is a time break.

l. Make any necessary stops and starts as indicated by time limits of sections within the test. Give additional directions if needed at these time breaks.

m. Stop the test at expiration of time and give directions for return of the testing materials. Where separate test booklets, answer sheets, supplementary testing materials, pencils, etc. have been issued, it is preferable to have each collected separately. Check the returns to be sure all materials issued are accounted for before dismissing the group.

Careful observance of the foregoing safeguards in testing will pay dividends by way of more accurate measurement.

4C. COORDINATION OF IDENTIFICATION, PERFORMANCE, AND WRITTEN TESTS.

1. The need for coordination of tests.

Since performance tests require more equipment and more personnel to administer than do identification tests or written tests, they may form a bottleneck in a testing program. With limited equipment and personnel, the time needed for completion of the performance test may be all that is available in the training program for testing. However, by proper scheduling, an entire class may be given

all three types of tests within the time limit that would be required to administer the performance test alone.

2. How coordination is accomplished.

The first step in coordinating the three tests into one test session is to answer certain questions about the performance test. How can the performance test be administered to the entire class within the time limits available? How many trainees can be tested at one time? How many groups will this make? How much time will be allowed for each group? The answers to these questions give a basis for planning the length and kind of performance test which can be given.

Having decided on the length of performance test that can be administered in the time allotted, the instructor can then build the written test and the identification test of such lengths so that the time required to administer them dovetails into the time limits dictated by the performance test. The scheduling of time periods and rooms is dependent upon the equipment available for the performance test and the personnel needed to administer the testing program.

An illustration typical of the general situation found in a training program will serve to indicate how coordination is accomplished.

a. *Situation.*
   (1) A three-hour testing period is scheduled.
   (2) A class of 48 trainees is to be given a written test, a performance test, and an identification test.
   (3) Fifteen instructors are available to administer the testing program.
   (4) The equipment available allows for testing twelve men at a time (4 each at 3 different stations) in the performance test.

b. *Procedure.*
   (1) The first question to be answered is "How many performance test sessions are needed?" Based on the situation above, it is necessary to schedule four separate sessions for the administration of the performance test. Twelve men can be tested at each session; therefore, forty-eight men can be tested during four successive sessions.
   (2) "How long shall each performance session be?" The first impression one is apt to get is that a three-hour testing period nicely divides itself into four forty-five minute periods. In practice, however, time must be allowed at the beginning of the testing session for giving general directions and additional time must be scheduled for moving from one examination room to another. Allowing five minutes at the beginning of the session for giving general directions and three, five-minute periods for changing rooms, four forty-minute periods are available for testing.
   (3) Build the performance test so that it can be administered in a forty-minute period, including time needed for specific directions.

(4) Build the identification test of such length that it can be given in one forty-minute period, including time needed for specific directions.

(5) Construct the written test of such length that it can be completed in a testing period of eighty-five minutes. (The time allotted for the written test is the sum of two forty-minute periods and the five minutes allowed between these periods.) Allow time to distribute, collect and check testing materials.

(6) Before the examination period, set up three rooms for the administration of the tests. These rooms should be close to one another to eliminate confusion in moving from one room to another and to conserve time in changing rooms.

(7) Assign instructors to their duties for the examination. The performance test requires thirteen instructors, one to act as supervisor and timekeeper and twelve to record the performance of the trainees. One instructor can administer the identification test and one instructor can supervise the written test.

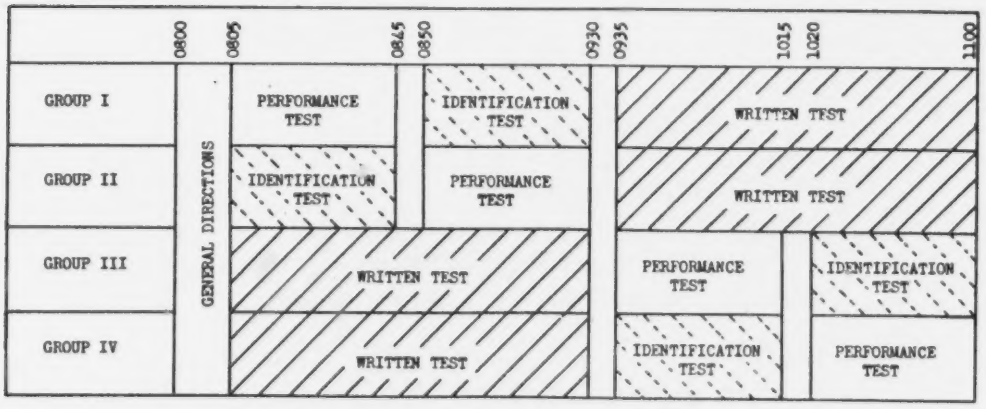


Figure 6.—Scheduling of performance, identification, and written tests.

(8) Assemble the class in the room where the written test is to be administered to give general directions.

(9) Divide the class into performance-test groups (in this example, four groups of twelve men each). Number these groups I, II, III, and IV.

(10) Send Group I to the room set up for the performance test. Send Group II to the room established for the identification test. Administer the written test to Groups III and IV.

(11) At the end of the first testing period of forty minutes, Groups I and II exchange places, Group I going to the identification test and Group II going to the performance test. Groups III and IV continue with the written test.

(12) At the end of the second testing period all four groups go to new testing rooms; Groups I and II return to take the written examination; Group III goes to the performance test; and Group IV goes to the identification test.

(13) At the conclusion of the third testing session of forty minutes Groups III and IV exchange places, Group III going to the identification test and Group IV going to the performance test. Groups I and II continue with the written test.

The diagram on the opposite page, Figure 6, indicates graphically the above scheduling for a morning session of three hours.

This procedure permits all the men to take all three tests within a period of three hours without conflict. Maximum use is made of the available testing time. Each group is permitted to finish each of the tests without interruption.

Some care must be exercised to prevent students from talking with each other about the tests while moving to and from the various examination rooms.

c. *Adaptations.*

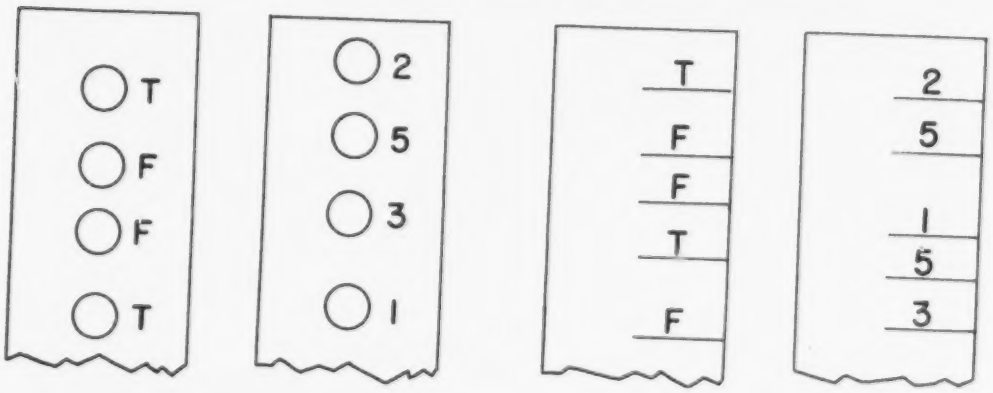
Where the situation requires setting up more than four performance test groups the scheduling becomes somewhat more complex. It is possible, however, to make up workable schedules by developing the written tests so that they can be administered in two or more sections, or by dividing the identification test into two or more parts.

CHAPTER V

SCORING TESTS AND GRADING STUDENTS

5A. MECHANISM OF SCORING

Care in setting up the form of the written test will ease the scoring burden after the test is given. Every type of question except the essay question can be put in a form that will permit the answer spaces to be placed in a straight column on the test page. In addition, the answer spaces can be kept clear of the written material which forms the test questions. These precautions will enable you to prepare and use the kind of scoring aids shown in Figure 7. Not only will this speed up the scoring process but scoring errors will be reduced and later analysis of the test will be made simpler.



Cut-out Scoring Stencils

On a strip of heavy paper punch out circles to fit the spaces used by students for their answers. Write the correct answer at the right (or left) of each hole in the stencil. Place stencil so answers show through the holes. Mark wrong answers with red pencil.

Strip Scoring Sheets

Mark off lines spaced to correspond with the answer lines on the test blanks. Write the correct answers on these lines. Line up answers on the scoring strip at the left of the answer column on the test papers. Check errors with red pencil.

Figure 7.—Scoring aids.

The cut-out scoring stencil is best adapted to scoring tests where the answers are single letters or numbers. The strip scoring sheets are more advantageous in scoring completion tests, where the answers are words or short phrases. At the top of each strip enter identifying information, such as the name of the test and the page number. At the bottom, place a note indicating the number of answers in the column. As each paper is checked, make a note at the bottom of the test paper showing both the number of right answers and the number of wrong answers for that column. If the scores are to be corrected for guessing, the number of omitted items should also be noted.

In scoring large numbers of papers both accuracy and speed are increased by checking one page on all papers, then the next page on all papers, and so on, rather than checking and scoring each student's complete test. Where several instructors are available the scoring task can be organized on an assembly line basis, each instructor checking an assigned page and the final man on the line adding up the page scores to obtain the total.

5B. USE OF THE SEPARATE ANSWER SHEET

A separate answer sheet may be used to good advantage for most tests. This arrangement, requiring the student to record all his answers on a separate paper instead of on the test itself, has a number of advantages. It will preserve the test form for repeated use since the student may be cautioned against making any marks on the test paper. In addition, having all the answers on a single paper rather than scattered through many pages of a test makes for simple, quick and accurate scoring. In Figure 8 a separate answer sheet is pictured, with answers marked in, preliminary to preparing one type of scoring stencil which may be used to correct the students' responses. The scoring stencil is shown in Figure 9.

The Separate Answer Sheet shown in Figure 8 is arranged for multiple choice, matching, and true-false items. Note that the multiple choice items have five spaces for marking answers. This corresponds to the five choices presented by each item. For the matching items seven answer spaces are provided. This is necessary when the number of possible answers in any one of the matching type units equals seven. If a greater number of possible answers exists for any matching unit, then additional answer spaces must be provided. The true-false arrangement is quite obvious. Space may also be provided on the separate answer sheet for writing in the answers to completion type items.

When the separate answer sheet is used, the directions to students should caution them against making any marks on the test papers. A brief explanation of the procedure to be followed in answering, together with one or two sample questions and answers, will insure that the students understand what they are to do. Nearly all of the students have had experience with separate answer sheets in previous tests and will require little instruction as to procedure.

SEPARATE ANSWER SHEET

NAME: John Jones DATE: 5 June 1947 TEST SCORE:

TEST: 20 mm INSTRUCTOR: Chief Williams CLASS: G-11

MULTIPLE CHOICE										MATCHING ITEMS								TRUE-FALSE		
(1)	X	2	3	4	5	(16)	1	2	3	4	X	(31)	1	2	3	4	X	6	7	(46)	T	X
(2)	1	2	X	4	5	(17)	1	X	3	4	5	(32)	1	2	3	4	5	6	X	(47)	X	F
(3)	1	X	3	4	5	(18)	X	2	3	4	5	(33)	1	2	X	4	5	6	7	(48)	X	F
(4)	1	2	3	4	X	(19)	1	2	3	X	5	(34)	1	2	3	X	5	6	7	(49)	T	X
(5)	X	2	3	4	5	(20)	1	2	X	4	5	(35)	1	X	3	4	5	6	7	(50)	T	X
(6)	1	2	X	4	5	(21)	1	2	3	4	X	(36)	1	2	3	4	5	X	7	(51)	X	F
(7)	1	2	X	4	5	(22)	X	2	3	4	5	(37)	X	2	3	4	5	6	7	(52)	T	X
(8)	1	X	3	4	5	(23)	1	X	3	4	5	(38)	1	2	X	4	5	6	7	(53)	X	F
(9)	1	2	3	4	X	(24)	X	2	3	4	5	(39)	X	2	3	4	5	6	7	(54)	X	F
(10)	1	2	3	X	5	(25)	1	2	X	4	5	(40)	1	X	3	4	5	6	7	(55)	X	F
(11)	1	X	3	4	5	(26)	1	2	3	X	5	(41)	1	2	3	X	5	6	7	(56)	T	X
(12)	1	2	3	X	5	(27)	1	X	3	4	5	(42)	1	2	X	4	5	6	7	(57)	X	F
(13)	1	2	3	X	5	(28)	1	2	3	4	X	(43)	1	2	3	4	5	6	X	(58)	T	X
(14)	X	2	3	4	5	(29)	X	2	3	4	5	(44)	1	2	3	4	X	6	7	(59)	T	X
(15)	1	2	X	4	5	(30)	1	2	3	X	5	(45)	1	2	3	4	5	X	7	(60)	X	F

Figure 8.—Separate answer sheet marked as key.

The Scoring Stencil shown in Figure 9 is merely a copy of the separate answer sheet with holes punched in the spaces corresponding to the correct answers. In the upper right and left corners of the stencil, holes have been punched to correspond with the two "x" marks appearing on the answer sheet. By lining up these holes in the stencil with the "x" marks on the answer sheets, the stencil will fall in its proper place over the answer sheet. The correct answer spaces will then show up through the holes in the stencil. Wrong answers are indicated by blank spaces in which the student's marking does not show through. These blank spaces should be checked in red by the scorer. If a large number of papers are to be scored it is advisable to mount the "keyed" answer sheet on a heavy piece of paper such as half of a file folder before punching out the stencil.

In using the scoring stencil it is preferable to *mark* the incorrect items, rather than only to count them. This will (1) insure greater accuracy in scoring, (2) make it easier to analyze the test at a later time, and (3) be useful in showing the individual trainee where he has made errors if the test results are followed up by class discussion or personal conferences.

SEPARATE ANSWER SHEET — Scoring Key

NAME: _____ DATE: _____ TEST SCORE: _____

TEST: _____ INSTRUCTOR: _____ CLASS: _____

MULTIPLE CHOICE					MATCHING ITEMS				TRUE-FALSE	
(1)	**1**	2	3	4	5	(16)	1	2	3	4	**5**	(31)	1	2	3	4	**5**	6	7	(46)	**T**
(2)	1	2	**3**	4	5	(17)	1	**2**	3	4	5	(32)	1	2	3	4	5	6	**7**	(47)	**T**
(3)	1	**2**	3	4	5	(18)	**1**	2	3	4	5	(33)	1	2	**3**	4	5	6	7	(48)	**F**
(4)	1	2	3	**4**	5	(19)	1	2	3	**4**	5	(34)	1	2	3	**4**	5	6	7	(49)	**T**
(5)	**1**	2	3	4	5	(20)	1	2	**3**	4	5	(35)	1	**2**	3	4	5	6	7	(50)	**F**
(6)	1	2	3	4	**5**	(21)	1	2	3	4	**5**	(36)	1	2	3	4	5	**6**	7	(51)	**F**
(7)	1	2	3	**4**	5	(22)	1	**2**	3	4	5	(37)	**1**	2	3	4	5	6	7	(52)	**T**
(8)	1	**2**	3	4	5	(23)	1	**2**	3	4	5	(38)	1	2	**3**	4	5	6	7	(53)	**F**
(9)	1	2	3	4	**5**	(24)	**1**	2	3	4	5	(39)	**1**	2	3	4	5	6	7	(54)	**F**
(10)	1	2	3	**4**	5	(25)	1	2	**3**	4	5	(40)	1	**2**	3	4	5	6	7	(55)	**F**
(11)	1	**2**	3	4	5	(26)	1	2	3	**4**	5	(41)	1	2	3	**4**	5	6	7	(56)	**T**
(12)	1	2	**3**	4	5	(27)	1	2	**3**	4	5	(42)	1	2	**3**	4	5	6	7	(57)	**F**
(13)	1	2	3	**4**	5	(28)	1	2	3	4	**5**	(43)	1	2	3	4	5	6	**7**	(58)	**T**
(14)	**1**	2	3	4	5	(29)	**1**	2	3	4	5	(44)	1	2	3	4	**5**	6	7	(59)	**T**
(15)	1	2	**3**	4	5	(30)	1	2	3	**4**	5	(45)	1	2	3	4	5	**6**	7	(60)	**F**

Figure 9.—Scoring stencil prepared from separate answer sheet scoring key.

It is advisable to "scan" or examine each student's answer sheet before applying the stencil in order that questions that have been answered in two places may be marked "incorrect." To insure that all such items are counted as incorrect draw a red line through all the possible choices for each item where two or more choices have been marked. Otherwise, students may select both the right and a wrong choice for a single question, and only the correct choice will show through on the stencil.

The student's score can be readily computed by adding up the number of errors shown by the red marks. Subtracting the number of errors from the total number of items on the test will give the student's score.

During the process of scoring a group of test papers it is a wise practice to check your accuracy by rescoring every tenth paper. In rechecking it is better to count the number of right answers if your method of scoring has been to count and subtract the number of wrong answers. Should the scoring of a test be turned over to someone else to do, it is well to run a random check on his scoring. One paper in every eight or ten should be rescored to see whether the scoring has been done properly.

When completion type items have been used in a test with separate answer sheets, preparation of the scoring stencil may be somewhat more complicated. The problem can be simplified by using separate strip scoring sheets for the completion type section of the test.

5C. CONVERTING TEST SCORES INTO STUDENT GRADES

1. Obtaining raw scores.

In determining a student's grade, it is first necessary to find the score he made on the test. In general practice, when the objective or short answer test is used, the score on the test is the sum of the correct answers. This sum may be obtained in one of two ways. The direct method is to count the number of right answers. The number obtained is the "raw" score. The indirect method is to count the number of incorrect items and the number of omitted items; these numbers are then subtracted from the total number of items on the test to obtain the raw score. Ordinarily the relative standing of the members of a class is directly related to the raw scores obtained.

In some situations, however, the scores obtained will place the students in inverse order. This is particularly likely to happen in certain types of performance tests where the scoring is done in terms of operations *not completed* or *incorrectly performed*, or in terms of time required to complete a task. In these cases it is important to remember that a low score indicates good performance and a high score poor performance. Where an examination includes separate tests, one or more of which yields inverse scores, the scores can not be directly combined. Suggestions for combining and weighting test scores will be found in Chapter VII.

2. The problem of guessing.

In multiple choice questions, and the other types in which the trainee is required only to choose the best of several given answers, a person who knows nothing about the subject is almost sure to get some of the right answers by pure chance. Thus, if a test were made up entirely of five-choice items, a person might get about one-fifth of the right answers, by pure guessing. But, where the group knows something about the subject being tested, the guesses are usually no longer "pure." The trainee will recognize that one or two or three of the choices is wrong, or inadequate, and will make his guess only between the remaining choices. Thus, he would get more of the right answers by guessing, but he deserves to get a better score because he knows at least enough to eliminate the worst answers even though he may not know the best ones.

a. *Correcting for guessing.*

There are standard formulas for correcting scores for guessing. These are based on the probability that, for every 2, 3, or 4 wrong answers given, one right answer was obtained by guessing. Thus, in a test made up of 120 items with four

choices per item, a trainee may have 90 right answers and 30 wrong answers. Since there were four choices on each item, he would guess wrong 3 times for every time he guessed right, or his *right guesses* will equal one third of his *wrong answers*. His corrected score, then, would be 90 (his right answers) minus 10 (one third of his wrong answers) or 80. This formula does not take into account that some of his wrong answers may be the result of wrong learning (instruction); it assumes that every wrong answer must have resulted from guessing rather than from "knowing too much that isn't so."

If a correction for guessing is to be made, separate counts of right and wrong answers or of wrong answers and omissions must be made, for obviously an omitted item is neither right nor guessed. If, in the example given above, the trainee had given 90 right answers, omitted 12, and given 18 wrong answers, his corrected score would have been 90 minus 6 (one-third of the *wrong answers*) or 84. In practical scoring the same result would be reached by the process of taking 120 (total number of items in the test) minus 12 (items omitted, no guess involved), minus 18 (items with wrong answers), minus 6 (correction for guessing), equals 84.

b. *Does not change trainees standings.*

Where all the trainees attempt all of the questions, applying the correction will result in lowering all scores, the lowest scores being lowered the most. Thus, the range between lowest and highest corrected scores will be greater than the range between lowest and highest "raw" scores, but each trainee will still keep his same relative standing in the group. Using again the 120-item test situation, Figure 10 illustrates how the correction increases the range but still keeps the scores in the same relative standing.

Right Answers	*Wrong Answers*	*Correction*	*Corrected Scores*
120	0	0	120
117	3	1	116
114	6	2	112
111	9	3	108
—	—	—	—
—	—	—	—
81	39	13	68
78	42	14	64
75	45	15	60
72	48	16	56
120 — 72 = 48, Range in uncorrected scores		120 — 56 = 64, Range in corrected scores	

Figure 10.—Table showing application of correction for guessing.

It is clear that in a situation like this there is nothing gained by the additional computation. Only when trainees are directed to omit questions rather than guess,

and when there are, in fact, a varying number of omitted items does the correction for guessing affect the relative standing of the trainees. For tests in which several different types of items are used, for example, 4-choice, 5-choice, matching, and completion, determination of a correction formula and computation of corrected scores are too complex to warrant the extra effort.

c. *Correcting for the true-false test.*

In the case of a true-false test, or a true-false section included in a test composed of several types of items, it is both easy and desirable to apply a correction for guessing. The element of chance in guessing is almost constant since there are only two possible choices. The trainee either recognizes that a given statement is true or that it is false, or he doesn't—and guesses. When he guesses his chance of guessing right equals his chance of guessing wrong. Therefore, for every wrong answer a right guess is assumed, and the corrected score is found by taking right answers minus wrong answers.

d. *Passing score should allow for guessing.*

The guessing element must be taken into account when setting the passing or "cutting" score on a test, and in setting up a system of converting test scores to grades or marks. The "cutting" score should certainly be somewhat higher than the score which could be "earned" by a person with no training. The problem in setting a cutting score is to determine the score which will be used in separating those who fail to qualify from those who have barely minimum qualifications. The person with barely minimum qualifications will answer some part of the items correctly because of what he knows; in addition he will "guess" the right answers to many additional items, either by knowing enough to eliminate some of the poorer choices in the items where he is not sure of his answers, or by pure guess.

Suppose a test is made up of 125 items in five-choice form. Presumably a person with no training could get a score of 25 right answers simply by marking answers 1, 2, 3, 4, and 5 at random, or by marking all the number 2 choices, or by following any pattern of marking the choices. This does not imply that 25 points should be subtracted from every score, for it must be assumed that the man who *knows* will not guess. Take a trainee who *knows* 40 of the items; he will make an informed guess or a "pure" guess on the remaining 85. If he makes "pure" guesses on all 85 he ought to get one-fifth right and four-fifths wrong. Thus he would have 17 additional right answers. If he made "informed guesses" for part of the items he would make a few more additional points. So the instructor's problem, in setting a cutting score, is to predict, as well as he can, the score that would be made by a barely qualined man. How many of the items would he be expected to know? On how many would he have to fall back on pure chance, a blind guess? The instructor's review of the test might turn up a table something like that shown in Figure 11.

Degree of Chance	*Items numbered*	*Number of Items*	*Contribution to Cutting Score*
Known (Chance is 1/1)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 19, 20, 21, 22, 23, 25, 26, 28, 30, 31, 33,	27	27
Guess from 2 choices (Chance is 1/2)	14, 18, 24, 29, 34, 35, 37, 38, 39, 41, 42, 43, 44, 45, 47, 48, 50, 51, 53, 56, 57, 58, 61, 71,	24	12
Guess from 3 choices (Chance is 1/3)	27, 36, 49, 52, 55, 62, 63, 64, 66, 67, 69, 70, 73, 74, 76, 77, 78, 81, 82, 83, 92, 95, 100, 104,	24	8
Guess from 4 choices (Chance is 1/4)	32, 40, 54, 59, 65, 72, 79, 80, 84, 85, 86, 89, 90, 91, 93, 94, 96, 97, 98, 101, 102, 103, 105,	23	6
Guess from 5 choices, Pure guess (Chance is 1/5)	46, 60, 68, 75, 87, 88, 99, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125.	27	5
		Expected Score — 58	

Figure 11.—Test performance of "Man with minimum qualifications."

This procedure can also be followed in setting up an expected top score for the test, and might yield a table such as that shown in Figure 12.

Degree of Chance	*Items numbered*	*Number of items*	*Contribution to Score*
Known (Chance is 1/1)	1-45, 47-59, 61-67, 69-74, 76-86, 89, 91, 92, 95, 97, 100, 103, 104,	90	90
Guess from 2 Choices (Chance is 1/2)	46, 60, 68, 75, 90, 93, 94, 96, 98, 101, 102, 105,	12	6
Guess from 3 Choices (Chance is 1/3)	87, 88, 107, 110, 112, 114, 116, 117, 119, 121,	10	3
Guess from 4 Choices (Chance is 1/4)	99, 108, 111, 115, 120, 122, 123, 124,	8	2
Pure guess (Chance is 1/5)	106, 109, 113, 118, 125	5	1
		Expected Score—102	

Figure 12.—Test Performance of "Best qualified man."

e. *Setting standards requires expert judgment.*

Clearly, the degree of chance assigned to each item in making tabulations such as these is only the instructor's estimate, both with respect to the value of each

item in showing qualification, and with respect to the way in which his "barely qualified" and "best qualified" men would respond to each item. But it is part of an instructor's job to make such judgments, to determine what kind of test performance is to be required in order to pass and in order to make a top grade. Every instructor and every leading petty officer is required to decide whether a man's performance merits a 2.5 or some other grade, and there is little but his experience and judgment to guide his decision in marking "proficiency in rate." On the old style written examinations the instructors had to decide, for each man's answers, whether he deserved to pass or what marks should be given, and the marks reflected a separate assessment of each man's qualifications, without reference to any standard. Whether the traditional Navy 4.0 scale or a "percentage" grading system was used, the instructor made a judgment of each paper's value in terms of the system. A mark of 2.5 doesn't represent 25/40ths of perfection; it is the grade given when the instructor believes a trainee is just good enough to pass with nothing to spare. Similarly a grade of 80% doesn't indicate that a man has learned 4/5ths of all there is to be known about a subject, but simply that his work meets the instructor's judgment of what a "good average" man can be expected to do. If there is a known relationship between scores on a test and performance on a job, statistical techniques can help to determine the passing score. Such known relationships to reliable performance standards are too often lacking. For most Navy situations, setting standards isn't a strictly mathematical or mechanical process; it requires the exercise of judgment as to what constitutes competence in any given field and how that competence can be shown. Once such judgment has been exercised, mathematical and mechanical devices can be applied in measuring, recording, and comparing, but the EXERCISE OF JUDGMENT IS FUNDAMENTAL.

f. *A short-cut method for setting passing scores.*

In section 5C2d a method of taking the guessing element into account when setting passing standards for objective type examinations was described. The method described there involves a review of each item by the instructor, to estimate the degree of knowledge and the degree of guessing to be expected of a man who barely deserves to pass. While such a review is desirable, situations may arise in which a short-cut "rule of thumb" method is needed. In such cases a reasonably fair cutting score can be estimated by going through the test and marking with a check-mark each item that the barely passing trainee is expected to *know* without any guessing—those items which represent the minimum essentials for doing the job. On the remaining unchecked items some degree of guessing can be expected, but the trainee will be likely to get more right answers than could be obtained by pure chance. If the number of items unchecked is divided by a number one less than the number of choices per item, a reasonably good allowance for guessing will be found. For example, an instructor prepares a test of 140 five-choice items and checks 47 which he thinks the barely passing student must know. There are then 93 items on which some guessing is expected, and

93 divided by 4, or 23, is the number of additional right answers expected as a result of guessing. So the estimated cutting score for the test would be found by adding the 47 "must know" items and the 23 expected right guesses, giving 70 right answers as the passing mark.

In case the test is made up of several sections in which different types of items are used, the same procedure can be applied to each section. Suppose a test is made up of 30 four-choice items, 60 five-choice items, and six matching units in which there are five questions and eight choices in the answer list. The cutting score could be estimated as follows:

	Types of Items			
	4-choice	*5-choice*	*Matching (5Q-8A)*	*Totals*
Number of items	30	60	30	120
Number checked ("must know")	18	28	16	62
Number unchecked	12	32	14	58
Divisor (chances minus 1)	3	4	7	
Expected right guesses	4	8	2	14
Items checked plus right guesses	22	36	18	76

The cutting score for the test would then be set at 76, the sum of the number of "must know" items and the number of expected guesses. Note that the factor of judgment is still present, but the instructor is not called upon to estimate trainee performance on every item in the test. If experience with the test itself shows that the estimated passing score is either too high or too low then the whole test should be re-examined.

3. Grading systems.

There are two systems of grading in common use in the Navy: (a) the 0-99 scale, and (b) the traditional Navy 0-4.0 scale. These systems are used in the grading of school work and job performance, and should not be confused with the Navy Standard Scores which are used in connection with the Basic Test Battery.

While these grading systems are referred to as 0-99 and 0-4.0 scales, there is really little justification for referring to the zero point as the base of either scale. Navy marks for proficiency in rating are rarely below 2.0, and school grades lower than 50 are rather infrequent. It would be more realistic perhaps to label the systems in terms of their two critical points, the "passing" grade and the grade given for best expected performance.

a. *School grades: The 0-99 or 63-99 scale.*

The Class A and Class P Schools under the cognizance of the Bureau of Naval Personnel have been directed to assign school grades on a numerical scale in which 63 is the lowest passing mark and 99 the highest mark given.

Navy instructors have often questioned why the highest grade should be 99 instead of 100 and have ventured the guess that it was used as a symbol that no

one is expected to be perfect. The actual explanation is much simpler; 99 takes only two columns when punched on an IBM card.

Since tests can be, and are, made up of any number of items, a method of changing test scores to school grades is needed. No matter how long or how short the test, whether it contains 10 or 217 items, if the scores are to be used in making up school grades some way must be found to translate the test scores into grades falling within the 0-99 range. Even if the test contained exactly 99 items, a translation would still be necessary since the number of right answers required for passing might be more or less than 63.

It is at this point that some very faulty thinking is often done. It is a common error to assume that the grade is obtained by figuring the percentage of correct answers on any test. On this theory a man who gets 40 right answers on a 50-item test would deserve a grade of 80; getting 150 right answers on a 200 item test would call for a grade of 75. The fault in this procedure is that it doesn't take into account the fact that tests differ in difficulty. A test might conceivably be so easy that getting 90 per cent of the answers correct would be a pretty poor showing. Or it might be so hard that getting as many as 50 per cent of the items correct would be a top-notch performance. Neither of these extreme cases is really typical of Navy tests, but it is true that there is wide variation in the difficulty of the tests used, wide variation in the difficulty of items included in any test (as there should be), and wide variation in the proportions of very easy, moderate, difficult, and very difficult items included in the tests.

If grading is to be fairly done these variations in difficulty must be taken into account. The percentage system of grading doesn't recognize either item difficulty or test difficulty. Applied to the test scores of a whole class or series of classes it gives the instructor an index of how difficult the test proves to be, but that is little help in the problem of converting test scores into grades.

b. *Navy Marks; the 0-4.0 or 2.5-4.0 Scale.*

The traditional Navy marking system uses a scale in which 4.0 is the best and 2.5 is the lowest passing score. It is customary under this system to grade individual papers using one or two figures beyond the decimal point and to extend averages to the third decimal.

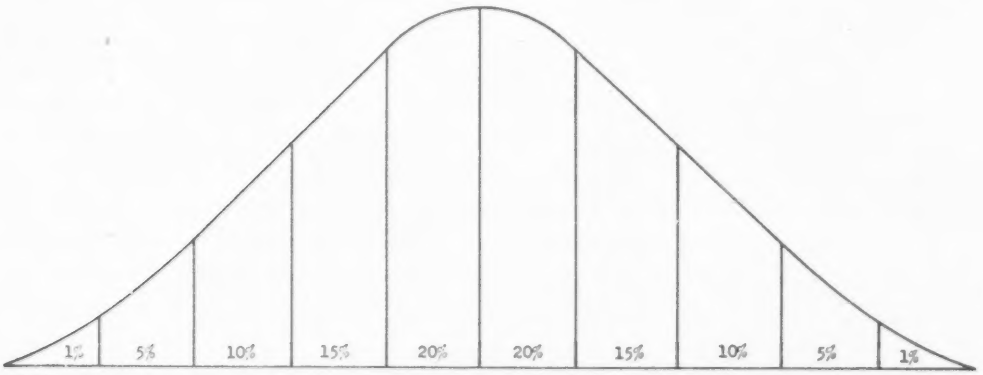
Translating test scores to the 4.0 scale by use of establishing percentage equivalents is just as faulty as using percentage as the basis of assigning grades on the 0-99 scale, and for the same reasons.

The Navy 4.0 marking system is essentially a rating scale rather than a grading system. As such the marks do not refer to any quantitative system at all, but are simply number symbols used to designate *quality* of performance.

The following is an excerpt from Article C-7821 of the Bureau of Naval Personnel Manual (Revised 1959 Edition) and is an example of a "Professional Performance" evaluation based on the concept of an average crew.

"To make the enlisted performance evaluation system thoroughly successful it would be desirable to assume that each command had an average crew assigned. The proportion of individuals who exceed the average in performance, or fall below the average in performance, then would be about the same for all commands. For the over-all good of the Navy and in the interest of the great majority of enlisted personnel, evaluations under the Enlisted Performance Evaluation System should be related in general to the average crew concept. Wherever this concept is strictly applicable, the distribution of marks in each separate trait for all personnel in a particular pay grade will approximate the optimum distribution of the "Professional Performance" evaluations illustrated in EXHIBIT A.

There must inevitably be a higher standard of required, as well as actual, performance with each higher pay grade. This results from the increasing experience level with each higher pay grade and from the fact that those individuals with less capabilities are eliminated by the competition for advancement. In view of this inherent increasing level of performance, it must be remembered that *individuals within a pay grade are to be evaluated solely against the performance of others in the same pay grade* and *not* against the performance of personnel in higher or lower grades. Under the Enlisted Performance Evaluation System the same percentage of senior petty officers will receive evaluations in each marking area for a particular trait as will personnel in pay grade E–3 or E–2. It is of course understood that the percentage of senior petty officers who merit the unsatisfactory evaluation in "Professional Performance" of "Inadequate" will be substantially less than the percentage shown in EXHIBIT A.



Extremely effective and reliable, works well on his own.		Highly effective and reliable. Needs only limited supervision.		Effective and reliable. Needs occasional supervision.		Adequate, but needs routine supervision		Inadequate. Needs constant supervision.	
(4.0)	(3.8)	(3.6)	(3.4)	(3.2)	(3.0)	(2.8)	(2.6)	(2.0)	(1.0)

Exhibit A

It is obviously impossible to take conduct into account in grading a test paper. Likewise, any single test is likely to be something less than a comprehensive, well balanced measure of all the factors which enter into competence in a given rating. A written test, for example, will measure some of the factors of knowledge and understanding of relationships, while it may require several different performance tests to measure the various skills which must be taken into account when the decision is finally made as to just what mark a man deserves on "proficiency in rate," or in progress in his training. But the mark given should be one which indicates the *quality* of his work, his degree of competence, with respect to whatever it is that is being measured by the particular test.

So the translation problem, essentially, is, "What score on this test represents a quality of performance that merits a 2.5 rating, meaning good enough to retain in the rate or in the training program?" Or, "What score represents performance that merits any of the other ratings as defined?"

c. *Navy Standard Scores.*

The Navy Standard Scores *are not used* as a means of grading achievement in schools or proficiency in rate. The system was developed as a means of reducing raw scores on various aptitude tests (the Navy Basic Test Battery, etc.) to comparable terms. The system is based on certain statistical assumptions concerning the distribution of scores made by large numbers of people picked out of a general population at random. Since the men in any given school, or the men in any given rate, have been picked for that school or rate on the basis of having certain defined characteristics, they do not form such a "random sample" of the general population, and therefore the assumptions basic to the Navy Standard system are not applicable in setting up a grading plan for a school or for marking proficiency on the job.

In the Navy Standard Score system the average is always placed at 50 and it is very unusual to find any scores higher than 80 or lower than 20. If the assumptions of the system are satisfied, a little more than two thirds of the people tested will have scores between 40 and 60, and only about one person in fifty will have a score lower than 30 or higher than 70. It should be clear that these scores are wholly different from the grades on the 0-99 scale used in the schools, where by far the greater part of the marks will be higher than 63. The only point of similarity between the two systems is that they both use two-digit numbers.

4. The translation graph.

There are many techniques which could be applied in setting up a procedure for translating test scores into student grades. (Some of those that have been devised and suggested are weird and wonderful.) But any grading plan has two fundamental purposes: (1) to establish a standard by which to determine whether the individuals graded meet, or fail to meet, the established qualifications, and (2) to provide a system of indicating how much better than barely passing each qualified man is, and how much each failure falls short of the qualification. The most critical point of the grading system is the dividing line between passing and failure. The

next most critical point is the effective top limit of the test, the best score that can be expected. With these two points established, it is clear that intermediate steps of quality can be laid out. Depending on the proportion of difficult material in the test, the intermediate steps could be made equal or unequal in terms of score points. Usually, unless the test is very badly balanced in difficulty, equal steps will provide a fair distribution of grades. The procedure described below provides a simple means of translating test scores to student grades:

STEP 1. CHOOSE A TEST SCORE WHICH WILL BE CONSIDERED EQUAL TO THE LOWEST PASSING GRADE. (63 on the 99 point scale, or 2.5 on the 4.0 scale.)

This step is the crucial one in the procedure and should be done without regard to the actual distribution of the raw scores on the test. It is at this point that the instructor must determine minimum standards in terms of the objective for which the training has been given. The question to be answered is "What is the minimum of knowledge and skills which the trainee must have in order to insure success on the job for which the training has been given? It is not "How many men should I pass and how many should I fail?" In setting this cutting score the instructor must consider the difficulty of the test itself. The point established for separating the passing from the failing will be higher for an easier test than that established for a more difficult one.

Having used his best judgment in setting the minimum passing score, the instructor may now check to see what happens when he applies this cutting point to the distribution of raw scores made by the students on the test. He may find that (1) the majority of scores fall below the cutting score with very few being in the passing group; (2) the majority of scores fall above the minimum passing score with few scores failing to meet the minimum; or (3) all scores are above the cutting point.

In case most scores fail to meet the minimum passing score, three things may have happened: (1) the men assigned to the training program may not have had the ability to successfully complete the training; they may have been potential "inapts" who should never have been assigned to the training program originally; (2) the training may have been inadequate in terms of the objectives of the course; or (3) the instructor's judgment in setting the cutting score may have resulted from faulty thinking.

Similarly, if all scores are above the passing point, (1) the trainees may have been well selected, (2) the instruction may have been excellent, or (3) the instructor actually may have set the cutting point too low.

Application of the cutting point to the distribution of raw scores *serves as a check* on the instructor's judgment; it *should not serve as a substitute* for his judgment. If, in a reevaluation of the situation, the instructor concludes that his original judgment was at fault, he may be justified in selecting a new minimum passing score. However, the decision must be made in terms of what standards are acceptable, rather than the number or percentage of trainees who will "fail" as a result of its application.

STEP II. CHOOSE A TEST SCORE WHICH WILL BE CONSIDERED EQUAL TO THE HIGHEST PASSING GRADE. (99 on the 99 point scale, or 4.0 on the traditional Navy scale.)

In a similar manner the instructor must use his best judgment in determining what raw score is worth the highest passing grade. Again the difficulty of the test needs to be considered before establishing this point. It is conceivable that in writing a test of 100 items, the instructor made it of such difficulty that even the best trained student could make a raw score of no more than 80. To assume that in order to obtain the highest possible grade (a 99 or a 4.0) an individual must answer all 100 items correctly is a fallacy. Knowing his own test, how difficult it is and how many questions even the best man might miss, the instructor would set the highest score which would be reasonably expected as the top grade of 99 or 4.0.

After the minimum passing and highest expected scores have been established, the rest of the process is mechanical.

STEP III. PREPARE THE TRANSLATION GRAPH.

In order to illustrate the mechanical process involved in the construction of the translation graph an example is presented, in Figure 13, for a class of 40 individuals who took an 80-item test.

Student	*Test Score*	*Grade*	*Student*	*Test Score*	*Grade*	*Student*	*Test Score*	*Grade*	*Student*	*Test Score*	*Grade*
Abel	60		Finch	60		Karp	28		Moon	46	
Adams	55		Fowler	72		Kris	57		Nacci	52	
Brown	52		Gans	48		Lamb	53		Netch	38	
Bruce	49		Gates	49		Laud	62		Noon	48	
Cady	42		Gill	51		Lord	68		Perry	63	
Camp	48		Gore	43		Lund	47		Pike	37	
Derry	32		Hale	35		Major	49		Pond	47	
Doakes	58		Hill	50		Mello	33		Ruch	53	
Downs	33		James	54		Mixer	48		Ryal	69	
Every	65		Judd	37		Moody	60		Wills	46	

Figure 13.—Sample class test record sheet.

On the basis of his best judgment, the instructor in this class set 74 as the score equal to the highest passing grade and a score of 35 as equal to the lowest passing grade. An inspection of the class record shows that four scores (33, 33, 32, and 28) fall below the minimum passing score of 35. The top score of the group is 72, two points below the test score set as the maximum likely to be made by any student. In this class, then, 36 of the 40 students, or 90 per cent of the group, made scores which were passing while four of the 40, or ten percent, fell below the minimum passing score. Notice that the cutting score was previously set on the basis of minimal standards and not on the assumption that ten per cent of the class should fail.

Procedures for preparing graph.

(1) On graph paper place *grades* on the base line at the bottom. Figure 14 illustrates the base line divided into ten equal units for conversion of the scores to the 0-99 point scale. Figure 15 shows the translation graph for converting to the 4.0 scale.

(2) On the vertical line at the left place the full range of *test scores* from the lowest possible score to the highest possible score, divided into equal units. For the test used in this discussion, 0 is the lowest possible score; and 80 is the highest possible score. The whole range from 0 to 80 is divided into 8 steps with 10 score points in each step.

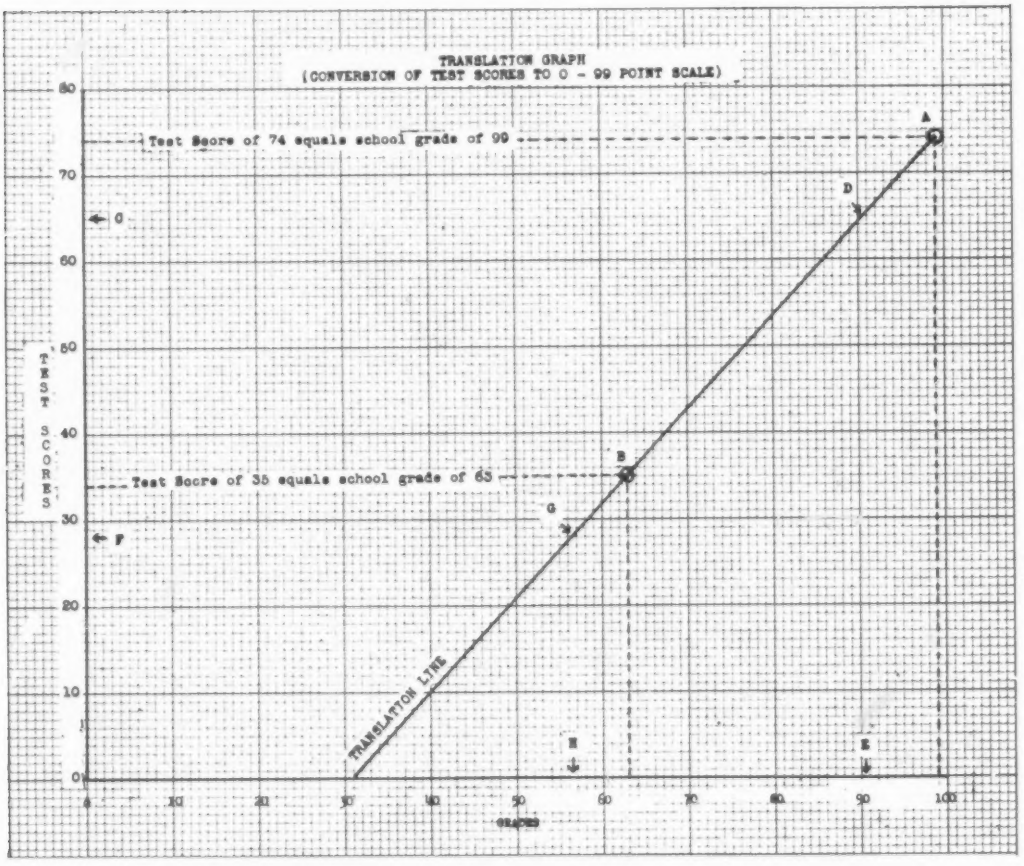


Figure 14.—Translation graph (conversion of test scores to 0-99 point scale).

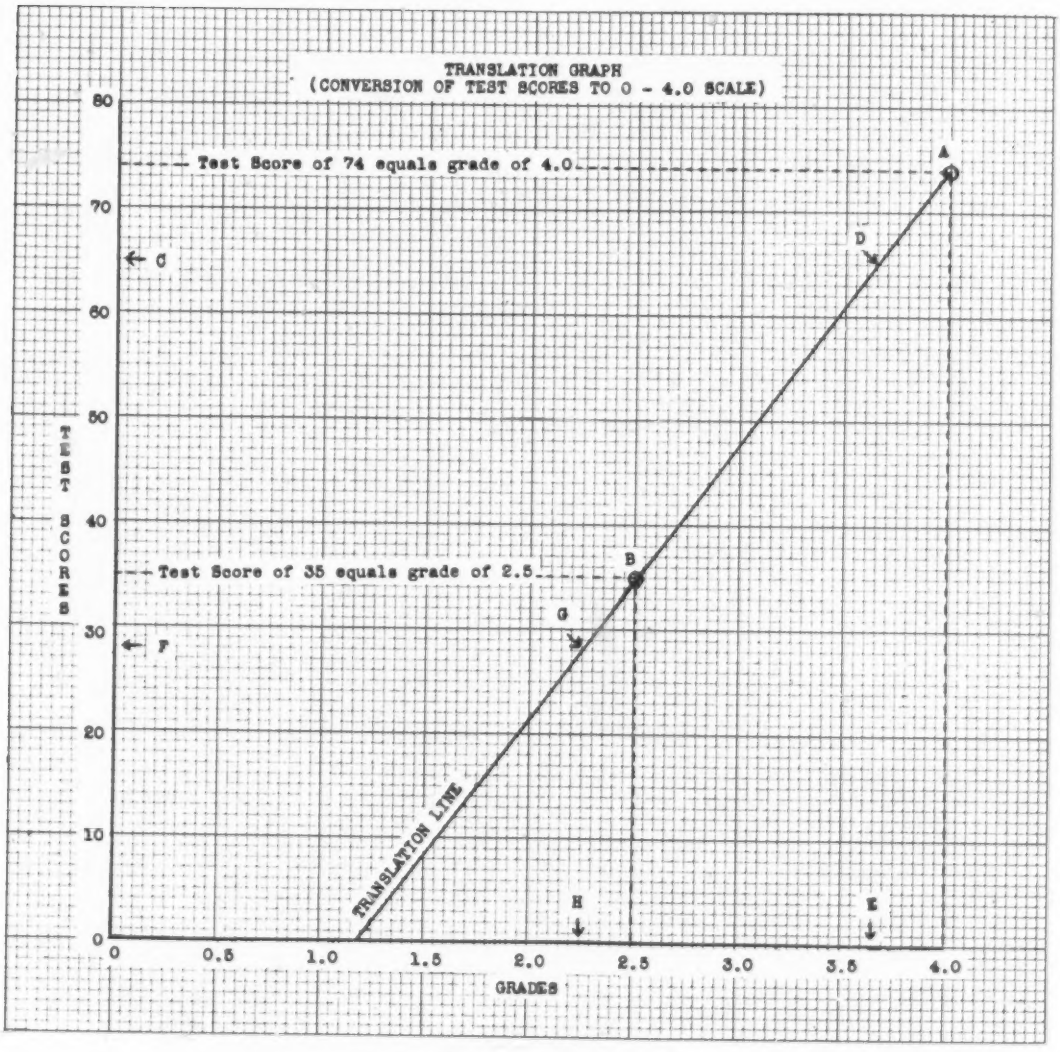


Figure 15.—Translation graph (conversion of test scores to 0-4.0 scale).

(3) Locate point "A". This point represents the *highest expected score* on the test and the *highest grade or mark* in the grading system. In the example used, 74 was judged to be the best expected score. Therefore, lay off a line beginning at the point representing 74 on the vertical scale and running across the paper to a point above the grade of 99 on the base line. Then lay off another line beginning at 99 on the base line and running straight up the paper to meet the first line drawn. Mark this point of intersection.

If the translation is to be made to the 4.0 scale, the second line will be drawn from 4.0 on the base line, straight up the paper to where it intersects the line representing test score 74.

(4) Locate point "B". This point represents the *lowest passing score* on the test and the *lowest passing grade or mark* in the grading system. In the example, 35 was established as the lowest passing score for the test. Lay off a line beginning at the point representing 35 on the vertical scale and running across the paper to a point above the grade of 63 on the base line. Then lay off another line beginning at 63 on the base line and running straight up the paper to meet the line drawn out from 35. Mark this intersection point.

If the translation is to be made on the 4.0 scale, this fourth line will be drawn from 2.5 on the base line, straight up the paper to where it intersects the line representing test score 35.

(5) Draw the slant line called the Translation Line. Draw a straight line connecting points "A" and "B" and extend it on down beyond point "B" to meet either the base line *or* the vertical line. In most cases the slant line will meet the base line, but if the test contains an unusually large number of very easy items it may meet the vertical line. Note that the line is *not* drawn to the 0-0 point; it is a *straight* line drawn from "A" through "B" and extended.

STEP IV. USING THE TRANSLATION GRAPH.

The translation graph may now be used to read the school grade for any test score in the class. Here are two examples. Follow them on each of the two graphs.

What is the grade of a man with a test score of 65? Find score 65 on the vertical line at the left (point C). Go straight across until you reach the Translation Line (point D). Go down until you reach the base or grade line (point E). The grade is 90.5 or 91 on the 0-99 scale or 3.65 on the 4.0 scale.

What is the grade of a man whose test score is 28? Find score 28 on the vertical line (point F). Go straight scross to the translation line (point G). Go straight down to the grade line and find the grade (point H). It is 56.5 or 57 on the 0-99 scale or 2.25 on the 4.0 scale.

Using this procedure will enable you to get the grade for any test score very quickly.

STEP V. BUILDING A CONVERSION TABLE.

In case the number of students taking a test is quite large, or in case the same test is used on subsequent classes or groups, it may be found profitable to use the translation graph to build a conversion table from which grades can be read directly. To do this, use the translation graph to determine the grade to be assigned to each test score from the highest expected score on down. Arrange these equivalent test scores and grades in tabular form and then use the table to assign grades. Figure 16 shows a portion of the conversion table computed from the translation graphs shown in Figure 14 and 15.

Test Scores	*Grades*		*Test Scores*	*Grades*		*Test Scores*	*Grades*	
	0-99 *Scale*	0-4.0 *Scale*		0-99 *Scale*	0-4.0 *Scale*		0-99 *Scale*	0-4.0 *Scale*
80	—	—	60	86	3.5	40	67	2.7
79	—	—	59	85	3.4	39	67	2.7
78	—	—	58	84	3.4	38	66	2.6
77	—	—	57	83	3.4	37	65	2.6
76	—	—	56	82	3.3	36	64	2.5
75	—	—	55	81	3.3	35	63	2.5
74	99	4.0	54	80	3.2	34	62	2.4
73	98	3.9	53	80	3.2	33	61	2.4
72	97	3.9	52	79	3.2	32	60	2.4
71	96	3.9	51	78	3.1	31	59	2.3
//	//	//	//	//	//	//	//	//
61	87	3.5	41	68	2.7	21	50	2.0

Figure 16.—Portion of a sample conversion table.

5. Giving Individual Grades.

In using either the translation graph or the conversion table to assign individual grades to trainee's papers, it is usually easier, and less likely to cause error, if the papers are arranged in order of their scores before the grades are assigned. Then, if five or six men have the same score it is only necessary to find the grade for that score once, and all who have the same score will be sure to get the same grade. Having the papers arranged in order of the scores is also desirable since it prepares them for other procedures which will be described in the next chapter.

CHAPTER VI

INTERPRETATION OF TEST RESULTS

6A. GETTING FULL VALUE FROM TESTING.

Why does a "spotter" report errors of range to the fire control room? To judge or grade the gun crew? This is certainly not the principal reason. The main purpose is to permit correction of range on the next salvo.

An achievement test should serve the same function as the "spotter" in gunfire. It should provide the data which can act as a basis for correcting weaknesses and shortcomings. There are practical and obvious reasons for grading students on their work, and a test is properly used in that connection. Yet a test which serves only that function hardly repays the instructor for the time, energy, and intelligence which go into its construction.

Below are four fundamental questions which must be raised following the administration of any achievement test.

1. What grades have the students earned?
2. Was this a fair test? How can it be improved for future use?
3. What weaknesses in knowledge or skill, calling for correction, are indicated by the test?
4. Are any of these shortcomings due to inadequate or poor instruction?

Note that getting a grade from the test is but one of the problems presented. A discussion of the methods to use in assigning grades from test scores was presented in Chapter V. Procedures designed to answer the three remaining basic questions are discussed in the paragraphs which follow.

6B. ANALYZING AREAS OF FAILURE.

The scores or grades earned on a test give only a general indication of what the trainees have achieved. But the test can serve also as a useful tool to determine whether instruction has been effective. If it is found that the trainees have fallen below expectations in particular phases of the program, then (1) those phases can be retaught, (2) the training program can be modified to give more attention to those phases in later classes, and (3) further study can be given to the methods of teaching used and to the suitability of the instructional materials and training aids. The test analysis procedures which are needed in determining whether teaching has been effective also lend themselves to determining whether

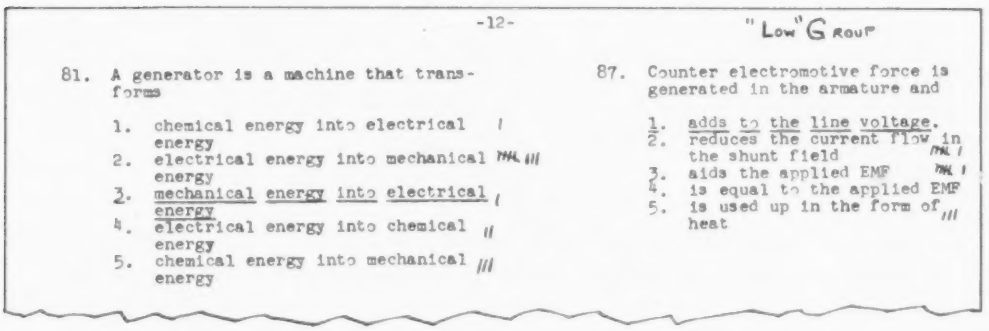
the questions used in the examination were effective as measures of achievement, and this is generally the first step taken. Obviously, it would be unwise to criticise the instruction program on the basis of test data which, itself, proved faulty.

The steps in the analysis described below apply to most written tests as well as to most performance tests. Analysis of essay questions and some types of performance records would require considerable adaptation of the procedure.

1. **Analysis of test items.**
   a. Arrange all test papers in order of excellence from the highest scores to the lowest scores.
   b. Take off one set of papers from the top or "high" group and another set from the bottom or "low" group. Just how many test papers to use for the "high" and "low" sets will depend on the number of students who took the test. In general, if 100 or more students were tested, a satisfactory division would be obtained by taking the upper and lower quarters from the total group of test papers. If less than 100 students had taken the test, it might be found more suitable to choose the upper and lower thirds from the total group.
   c. Now it is necessary to find out how well the students in both the "high" and the "low" groups answered each question. It is reasonable to assume that the "high" group will represent the better students in the class while the "low" group will represent the less able students. As will be shown later, an analysis of how these two groups answered each question gives valuable information as to the test, the students, and even the teaching of the subject. To obtain such information it is necessary to make a separate tally of the actual answers given by the "high" and the "low" groups to each question. This can be done quite conveniently by taking two blank test papers and recording the answers given by the "high" group on one of the test blanks, and the answers given by the "low" group on the other test blank. This means making an actual tally of the way every student answered each of the questions. To make this procedure clear, tallies of the answers for a "high" and a "low" group on just two questions of a test are presented in Figure 17 which follows. In this illustration 45 students had taken the test, and the "high" and "low" groups are composed of 15 students from the top and 15 from the bottom of the total group tested. Note how the tally marks indicate the number of men selecting each choice on both questions. Note that certain choices are underlined to indicate the correct answers. Note also that the final tabulation for both groups has been placed in columns on the test paper used for the "high" group. This will make for ease in comparing the results for the two groups.

2. **Interpreting the tallies.**

What information can you extract from these facts? To a large extent the analysis will assist in answering the following questions.



-12-

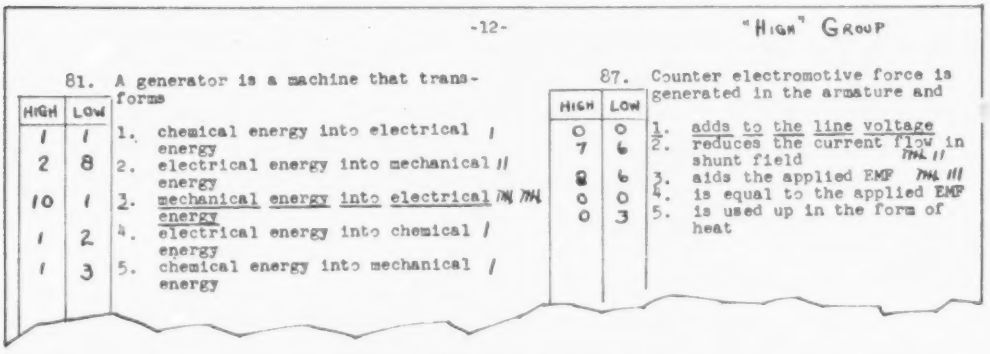
"Low" Group

81. A generator is a machine that transforms

1. chemical energy into electrical energy /
2. electrical energy into mechanical energy ||| |||
3. mechanical energy into electrical energy /
4. electrical energy into chemical energy ||
5. chemical energy into mechanical energy |||

87. Counter electromotive force is generated in the armature and

1. adds to the line voltage.
2. reduces the current flow in the shunt field |||| /
3. aids the applied EMF |||| /
4. is equal to the applied EMF
5. is used up in the form of heat |||



-12-

"High" Group

81. A generator is a machine that transforms

HIGH	LOW	
1	1	1. chemical energy into electrical energy /
2	8	2. electrical energy into mechanical energy //
10	1	3. mechanical energy into electrical energy |||| ||||
1	2	4. electrical energy into chemical energy /
1	3	5. chemical energy into mechanical energy /

87. Counter electromotive force is generated in the armature and

HIGH	LOW	
0	0	1. adds to the line voltage
7	6	2. reduces the current flow in shunt field |||| ||
8	6	3. aids the applied EMF |||| |||
0	0	4. is equal to the applied EMF
0	3	5. is used up in the form of heat

Figure 17.—Tallying of answers.

a. *Is The Test Item Any Good?*

In the illustration given, question 87 was missed by every man in both "high" and "low" groups. There's a danger signal. Either the question is very difficult indeed, or else something is radically wrong with the question. Actually, the item is at fault. Choice 1, which is indicated as the correct choice, is actually incorrect. Instead of the phrase "adds to the line voltage," it should read "opposes the line voltage." This is probably a careless slip on the part of the test maker. Its presence is immediately indicated by observing the tally.

Whenever a supposedly *wrong choice* is selected by a high percentage of the students in both the "high" and the "low" groups, the item should be examined carefully. Perhaps inadvertently this "wrong" choice has been made into a correct answer, or the *best* of the choices offered.

Likewise, if students in the "low" group give the correct answer to an item more frequently than students in the "high" group this item should be given close scrutiny. It may not necessarily be an incorrect or weak item, but the fact that the poorer group of students made a better score on the item points in the direction of some deficiency. It would be well to attempt to discover which choice attracted the good students. This may give a lead as to the possible deficiency in the question, or in the way the topic was taught.

Another weakness that will be shown up by the tallies is the item which all or nearly all of the trainees answer correctly. There should be a few such

items in every test, but sometimes an item proves "easier" than there is any apparent reason to expect. When this happens it may be caused by having all of the wrong choices so clearly incorrect that anyone with the slightest knowledge of the topic could eliminate them, or by having tell-tale errors in the way the wrong answers are written.

b. *In What Areas Are the Students Weak?*

Note that only 11 out of 30 students answered item 81 in the illustration correctly. Nearly two-thirds of the students failed the item. Here is a definite weakness on an important concept. Some corrective action is called for and should be taken. However, this concerns only a single question. In actual practice, you would group your test items by major areas. Then your tallies would show strength or weakness in a whole group of questions within an area of study. This is far more essential than just reviewing the test and correcting wrong answers. If a large section of the class is notably weak on a whole series of questions related to some broad aspect of the subject, clarification of the entire area might be needed. Merely dealing with answers to specific test questions would be inadequate.

c. *Are Any Teaching Weaknesses Revealed?*

Not every shortcoming in learning can be attributed to student failure. It is conceivable that poor teaching may play its role. Can this be said concerning either question in the preceding illustration? Question 87 has been rejected because of the slip in connection with framing the correct choice. It can tell us nothing about how the concept involved was taught since no information can be obtained on how it would have been answered if the question were correctly phrased. But item 81 does tell a story. Two-thirds of the good students answered this correctly. For the better students, the concept was taught sufficiently well to impress a substantial majority. The less able students failed miserably. Only one answered correctly. Eight of the fifteen have confused a generator with a motor. This is not necessarily a reflection on the teaching, but it surely indicates the need for a careful, simplified explanation of the principle involved so that the slower students can grasp all the elements of this problem.

For the moment suppose that only two students in each group, "high" and "low," had answered Question 81 correctly. Assume further that this same weakness appeared on most of the questions dealing with the subject of generators, though on other subjects at least the "high" group had done rather well. The inference would be reasonably strong that the teaching on generators had somehow fallen short. Upon further study of the test and of the teaching of the subject of generators, other factors may be found which contributed to this general weakness. The really important thing, however, is that the analysis of the test raises issues, requires self-examination by the instructor, and opens the way to improvement of the educational program.

6C. MAKING BETTER TESTS

In the schools which have a new class coming along each week or month, or some other regular interval, and in commands where advancement in rating examinations are given every quarter, it is a good practice to build up and maintain a stockpile of testing material. Keeping files of tests that have been used is one way of doing this, but it will save time in the long run and make for constant improvement in the tests prepared if a systematic program of stockpiling is followed.

1. Prepare the Item Cards

The first step in developing a stockpile is to place each test item or unit on a card. While 3″ by 5″ cards are large enough for most multiple-choice items, a larger card may be preferable since there may be matching units, or units built around an illustration or a diagram, which will require more space. Besides, space should be provided for index notations and the record of the tests in which the item is used and how effective it proves. Cards 5″ by 8″ are a good size for most purposes.

In addition to the item itself there should be index notes to make it easier to maintain the filing system. When an item is first added to the file it is desirable to make notations as to:

a. The school or rating group for which the item was designed.
b. The topic or examination subject to which the item belongs.
c. The stage or phase of the training program at which the item is most applicable, or the classes within a rating for which the item is most appropriate.
d. The source from which the item content is taken, the authority for the "right" answer; for example, BJM p ____, BuSandA Manual Sec. ____, ALNAV ____, Regs. Art. ____, etc.
e. The date when the item was prepared. Dating helps to indicate items which become obsolete through changes in Bureau Manuals, Technical Instructions, OrdAlts, etc.
f. The person who prepared the item and was responsible for verifying the right answer.
g. Whether the item was an original or a "steal" from some other publication, school, or command.

2. Set Up a Filing System

Having the items on cards is of little use unless they are classified as to subject matter. The filing system should be set up to fit the particular situation. In a school the major topics in the school curriculum may provide the basis for organizing the file. Advancement in rating item files may be classified according to the rating and examination subjects within the rating. Since the same examination subjects occur in several ratings it may be preferable to organize the file

primarily on the examination subjects. If the file is organized primarily on the ratings, some cross reference system may be needed to avoid making a lot of duplicate cards. The cross-reference cards can be flagged, or a different color card can be used.

3. Keep a Record of When and Where the Item Was Used and How It Performed

When the same item appears time after time in successive tests the word gets around. There are few topics, worth testing, for which it is impossible to provide some variety in the questions asked, the "decoy" choices provided, or the statement of the problem. An item that has been used in one test should be "given a rest," but may be used again after one or two or three alternates which get at the same idea have been used. In an advancement in rating examination program, if the same examination subject is listed for two or more rates, it is preferable to use an item in two or more tests given at the same time, rather than use it in one of the rates one quarter and in another the next. Each time an item is used a notation should be made on the back of the card showing the test in which it was used and the date on which the test was given. These notations can be made very brief. For example, "SK2, 4/15-47, #16" would indicate that the item was used in the test for advancement to Storekeeper Second given on 15 April 1947 as question number 16; "6wk, 6/13-47, #24" would mean that the item was used in the test given at the end of the sixth week of training in a school, on 13 June 1947, as question number 24. This kind of data is the minimum record. Item analysis data such as may be obtained by the procedures described earlier in this chapter should be added whenever it can be obtained. Unfortunately, the number of men taking advancement in rating examinations at any one time is usually too small to provide for comparison between the performance of "high" and "low" groups. However, it would be useful to have a record, even for small numbers, of the numbers who gave the right answer among those who passed and among those who failed. Item analysis data should be entered on the same line as the notation of the date of use and test in which used, in order to avoid confusion. Thus a full notation might appear as: "SK2, 4/15-47, #16; N-7, 1/2F - 3/5P" indicating that, when used in the advancement test for Storekeeper Second on 15 April, seven men took the test; one of the two who failed got the right answer and three of the five who passed gave the right answer.

Or, in the case of a school exam the notation might run: "6wk, 6/13-47, #24; N-63, H-L 20; (1) 6-4, (2) 1-6, (3)* 12-3, (4) 1-5, (5) 0-2" indicating that, when used in the sixth week test, given 13 June 1947, 63 men took the test and the analysis was made on the "high" and "low" twenty papers; six "high" and four "low" men chose answer 1, one "high" and six "low" chose answer 2, twelve "high" and three "low" chose the correct answer, 3, one "high" and five "low" chose answer 4, and none of the "high" and two of the "low" chose answer 5.

4. Maintain the File

The test analysis data may show that an item needs revision. If the basic item remains the same, for example the revision consists of changing one of the "decoy" answers which doesn't attract, or one which is too nearly correct, the change can be entered on the original card and dated. If the item proves too difficult at one level, the index notation of where it can be used can be changed. If an item proves ineffective, or analysis shows that there are major defects, a new item should be drafted and the ineffective one discarded.

Changes in the basic material from which test items have been developed may make some items obsolete. For example, an item requiring the computation of the monthly rate of pay of a petty officer first class with eight years service while on sea duty would be made obsolete by a change in the base pay. Substitute items should be prepared in accordance with the new instructions and the old ones discarded. Having the cards classified and indexed makes it easier to locate material which needs revision when such changes are published. The changes should be made immediately upon the receipt of new instructions, since there is danger that an obsolete item will be picked from the file when an examination is being prepared, without rechecking whether the content is in accord with current directives.

CHAPTER VII

WEIGHTING AND COMBINING TEST SCORES

7A. THE NEED FOR WEIGHTING SCORES

1. Grading Trainees Requires Combination of Measures

When final grades are assigned in a school, or when an examination grade is based upon a combination of several tests such as a written test, an identification test, and a performance test, or upon a series of tests in different subjects, each of the measures used contributes some part in determining the relative standing of the members of the group being graded. Just what part is contributed by each of the measures may be accidental, unless some care is taken to assure that each measure contributes in accordance with the value planned for it. *There is no such thing as an unweighted combination.* The weighting is either accidental or planned. Most instructors have a pretty clear idea of the relative importance which should be assigned to each test or measure that they use, and they frequently plan to assign weights to various scores, but these assigned or *"apparent"* weights may turn out to be quite different from the *effective* contribution made by each measure.

2. Effects of Accidental Weighting

A rather simple illustration will show the kind of results which may come from accidental weighting of test scores. An instructor has given two tests, one of which he considers to be important, the other relatively unimportant. The more important test included 100 items and the scores ranged from 70 to 90. The less important test included 60 items and the scores ranged from 10 to 50. Since the more important test was nearly twice as long as the less important, and since the average score was more than twice as great, the instructor decided to combine the scores directly. Now, assume that the man who got the highest score on the more important test got the lowest score on the other, while the man with the lowest score on the more important test got the highest score on the other. Clearly, the man who had the highest score on the more important test ought to outrank the man who had the lowest score on it, if the tests were properly weighted in the combination. But in this case high score on the important test (90) plus low

score on the unimportant test (10) yields a combined score of 100, while low score on the important test (70) plus high score on the unimportant test (50) yields a combined score of 120. The *effective* weight of the relatively unimportant test was great enough to reverse the positions of the high and low men on the more important measure when the scores were combined. While it *appeared* that the longer, more important test was being given twice the weight of the shorter test, the *effective* weight of the shorter test was actually much greater than the weight of the longer test.

3. How Does It Happen?

The error made in this situation was in assuming that the length of the tests, or the numerical value of the averages on the tests, would determine their relative contribution to the final relative standing when the scores were combined. The key to the situation, in the example given, lies in the *differences* between the highest and lowest scores on each of the two tests. Note that on the longer test there was a difference of only 20 score points between the highest and lowest scores; while there was a difference of 40 score points on the shorter, less important test. The *effective* weight of each measure that is used in a combination depends upon the relative *variance* of the measures, not upon the length of the tests nor upon the size of the average scores made on them.

Variance is a technical term used in statistics. Computation of variance, while not difficult, is somewhat complex and tedious. For practical purposes other measures of the way in which test scores are distributed can be used. The procedure which will be described in this chapter used the most easily found index, the range. Other measures which could be used include average deviation, standard deviation, and semi-interquartile range. But these are also tools of the statistician, and this discussion is intended only to provide a simple, workable, and useful procedure which will help the instructor to use his test scores fairly and with approximately the weight that he intends.

To illustrate how the range affects the weight of a test when combined with others, take the following situation. A group of 15 trainees has taken three different tests. On test "A" there were 50 items and the scores ranged from 30 to 43, averaging 36; on test "B" there were 90 items and the scores ranged from 60 to 88, averaging 72; on test "C" there were 100 items and the scores ranged from 75 to 88, averaging 81. The following table, Figure 18, shows how the individual members of the class ranked on each of the tests, their ranking on the combined scores, and what their ranking would have been if the scores had been combined so as to give the tests equal *effective* weight.

Trainee	Test "A"		Test "B"		Test "C"		Combined		Rank if weighted equally
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	
a	41	2	88	1	78	13	207	1½	3
b	43	1	80	2	84	3	207	1½	1
c	38	3½	75	4	85	2	198	3	2
d	37	5½	77	3	82	5½	196	4	5
e	38	3½	74	5½	81	8	193	5½	6
f	37	5½	73	7	83	4	193	5½	4
g	30	15	72	8	88	1	190	7	7
h	36	8	74	5½	77	14	187	8½	10
i	36	8	71	9	80	10	187	8½	9
j	35	10	69	11	82	5½	186	10	8
k	34	11½	70	10	79	11½	183	11	12
l	32	14	67	12	81	8	180	12	13
m	34	11½	64	14	81	8	179	13	11
n	36	8	66	13	75	15	177	14	14
o	33	13	60	15	79	11½	172	15	15
Highest	43		88		88		207 (Sum of highest 219)		
Lowest	30		60		75		172 (Sum of lowest 165)		
Range	13		28		13		35 (Sum of ranges 54)		
Average	36		72		81		189		

Figure 18.—Effect of combining scores on tests with different ranges.

There are several points worth noting in this table. First, individual standings on test "B" are more nearly comparable to the standings on the combined scores than are those on either of the other two tests. Second, the standings on test "A" are more nearly comparable to both the standings on combined scores and those on test "B" than are the standings on test "C". Third, the individual standings obtained by simply adding the "unweighted" scores differ from those obtained when equal *effective* weights are applied. (The "Rank if weighted equally" column in the table was computed on the basis of scores weighted to equate variance.) Fourth, the test with the greatest range (difference between high and low scores) had the greatest effective weight in determining the standings on combined scores; although the ranges of the other two tests were equal, test "A" had a greater effective weight than test "C" because the standings on test "A" tended to correspond with those on test "B". (Correlation between "A" and "B" was higher than between "C" and "B"; the effective weight of a test is increased if it is related to another measure used in the combination.)

It may also be noted that the sum of the highest scores made on the three tests (219) by various trainees was greater than the highest combined score made by any one trainee, while the sum of the lowest scores made on the three tests (165) by various trainees was lower than the lowest combined score made by any one trainee. Consequently the sum of the ranges on the three tests is greater than the range of the combined scores. On the other hand, the average of the combined scores is the same as the sum of the averages on the three tests. The greater the

number of measures included in a combination, the more this restriction in combined ranges will become apparent, especially if there is little relationship between the measures.

Anyone who may want to play with the problem a little further can prove to himself that the number of items in the test, or the average number of right answers *does not affect* the weight of each test in the combination. If the lowest score made on each test were subtracted from each of the individual scores, then the new scores on Test "A" would range from 0 to 13, on Test "B" from 0 to 28, and on Test "C" from 0 to 13. The individual combined scores would range from 72 to 42. The individual standings on each of the tests and on the combined scores would remain exactly the same.

4. Using the Range in Estimating Weights

As pointed out earlier, the range is not the most accurate index to use in determining the effective weight of a test in a combination. However, it is easily found, and weights determined by using comparative ranges as a guide will approximate those which the instructor intends to apply. If the difference between the highest and lowest scores on a test is 20, it means that the effective measuring range of the test is divided into 20 scoring units; if the difference in scores is 40, the effective measuring range is divided into 40 scoring units. When the scores on these two tests are added, the one having the range of 40 units will have twice the effect of the test with a range of 20 units. If the two tests are to be given equal effective weight in determining individual standings, the number of units in the ranges will have to be made equal, which can be done in this case simply by multiplying the scores on the first test by two.

The general principles in weighting scores on tests when combining the scores are

a. The weight of the measure varies with the number of score points in the range.

b. Equal effective weights for combined measures can be obtained by making the ranges equal.

c. Desired effective weights can be obtained by making the ranges proportional to the desired weights.

The first of these principles has already been illustrated in section 3. If a combination in which each of these tests is given equal weight is desired, the ranges could be made equal by multiplying each score on Test "A" and on Test "C" by 28/13ths, or by multiplying each score on Test "B" by 13/28ths. But approximately the same result will be obtained by multiplying each score on Tests "A" and "C" by two. In the table below, Figure 19, this is done.

Trainee	*Test "A" Score x 2*	*Test "B" Score (1)*	*Test "C" Score x 2*	*Total Score*	*Rank on Total*	*Rank if weighted equally*
a	82	88	156	326	2	3
b	86	80	168	334	1	1
c	76	75	170	321	3	2
d	74	77	164	315	4	5
e	76	74	162	312	6	6
f	74	73	166	313	5	4
g	60	72	176	308	7	7
h	72	74	154	300	10	10
i	72	71	160	303	8½	9
j	70	69	164	303	8½	8
k	68	70	158	296	11	12
l	64	67	162	293	13	13
m	68	64	162	294	12	11
n	72	66	150	288	14	14
o	66	60	158	284	15	15

Figure 19.—Effect of combining scores on tests with ranges approximately equated

It will be noted that the individual standings obtained by this approximate weighting correspond very closely to the standings obtained when equal effective weights determined by the variance of the tests were applied. Seven of the fifteen trainees were given exactly the same rank in class and none was displaced more than one number in rank.

7B. PROCEDURE FOR WEIGHTING

The process of combining scores on tests or marks on performance is made easier if a definite system is established and followed. The steps outlined below and the illustrations of how they can be applied will serve as guides in setting up a system suited to your situation.

1. General Procedure

The steps leading from raw scores on a number of sets of tests to individual grades based on weighted combinations of the test scores are as follows:

a. List the different measures that are to be included in the combined grade or "final multiple."

b. Decide on the relative weight that each test or measure *should* have. For practical purposes this is best expressed in terms of percentages of the total.

c. Find the range of each test or measure.

d. From the relative size of the ranges, calculate the weight which each measure would have if the scores were simply added together. (Find the sum of the ranges and divide each range by that sum to find the percentage contribution.)

e. Determine the number by which the scores on each test must be multiplied in order to change the ranges so that the contribution of each test to the

total grade will be about what you decided it should be, in step b. above. (Divide the percentage set up in step b. by the percentage calculated in step d.) This number is called the multiplier.

f. Multiply the individual scores on each test by the proper multiplier.

g. Add these "weighted" scores for each individual to get his total score or final multiple.

h. Make a translation graph for these total scores and determine the grade on the 99-point scale or mark on the 4.0 scale which each trainee has earned.

2. Detailed Procedure for Weighting

The separate steps by which this combining process may be carried out are listed below. There are two major phases of the procedure: (1) determining the multipliers, and (2) applying the multipliers to the individual scores. The accompanying table, Figure 20, illustrates the steps required to obtain the multipliers for each test.

a. Set up a table.

Make up a tabulation form similar to that in Figure 20. The significance of each of the column headings will be shown as the steps in the procedure are described below.

b. Column 1, Test.

Write the name or other identifying data of each test or measure to be included in the combination. No special order need be followed.

c. Column 2, Desired Contribution.

Decide the percentage of the final grade which each measure should contribute. In the example, scores on six tests are being combined; Test C has been considered most important and is assigned 30 percent as a desired contribution. Test B has been considered least important and is assigned 8 percent as a desired contribution. Since this step is very important, several persons who are well acquainted with the requirements to be graded and with the tests that have been used should be consulted. In making this decision as to the relative importance of the tests the following factors should be considered:

(1) *The significance of the content of the test.* What abilities or qualities are measured by the test? How important are these things for the general ability which the total grade is supposed to represent?

(2) *Amount of curriculum covered.* In school situations, the time spent on each topic in the curriculum can serve as a rough guide to their relative importance. Other things being equal, a test covering ten periods of instruction is twice as important as a test on five periods of instruction.

(3) *Overlapping between the tests.* In any set of tests covering a program of training there is likely to be considerable overlapping among the tests, but perhaps two or three tests will be much more highly related to each other than to the remaining tests of a series. Such inter-rela-

1 Test	*2 Desired Contribution (%)*	*3 Lowest Score Made*	*4 Highest Score Made*	*5 Range (4-3)*	*6 Contribution if Added (%)*	*7 Theoretical Multiplier (2/6)*	*8 Practical Multiplier*
A	12	61	93	32	14	0.9	1
B	8	10	36	26	12	0.7	1
C	30	73	97	24	11	2.7	3
D	20	40	68	28	13	1.5	1½
E	15	70	88	18	8	1.9	2
F	15	25	120	95	43	0.3	½
Totals	100			223	101		

Figure 20.—Table used in determining multipliers to be used in weighting and combining scores on several tests.

tionship may be caused by similiarity of the material covered, or by similarity in the type or form of the tests. The tests which are most highly interrelated will tend to contribute more to the total score than the size of their relative ranges would indicate. To compensate for this tendency, the "desired contribution" for each of the highly related measures should be reduced.

(4) *The accuracy of the test.* A long test that covers its subject thoroughly is likely to be a more accurate measure than a short one, and should be assigned a larger contribution in determining final grades. In deciding this factor the length of time required to take the test and the type of the test should be considered as well as the number of items included. Further, a test that has been poorly administered, with various distractions or under improper proctoring, or one which includes scoring errors is a relatively inaccurate measure. The contribution of such a measure in determining final grades should be reduced, or perhaps eliminated.

(5) *Check the total.* The sum of the desired contributions should add up to 100 percent.

d. Column 3, Lowest Score Made.

For each test or measure, write in the lowest score that has actually been made by any individual. This is, of course, the first step in determining the range of the test. Occasionally there will be cases in which the lowest actual score should not be used. If a single score far below any others appears on

a given test, it is likely to have been caused by some peculiar circumstance affecting the individual who made it rather than the entire group. For example, a man who had just had his eyes dilated for examination before taking the test would be at a disadvantage, as would one who had just returned from a week in sick bay. In such instances the next lowest actual score may be used.

e. Column 4, Highest Score Made.

Similarly, write in the highest score that has actually been made by any individual on each measure. Again there may sometimes be reason for excluding an extremely high score. For example, one of the men may have taken the test a day or so later than the others and had the advice of the others before taking it.

f. Column 5, Range.

This is found by subtracting the lowest score, entered in column 3, from the highest score, entered in column 4, for each test. After the range of each test has been found, *find the sum of the ranges,* and write this sum in the TOTALS space at the bottom of the column. Check your addition to be sure that is correct.

g. Column 6, Contribution if Added.

This shows the effective weight that each test would have in the final grade if the total scores were obtained simply by adding each individual's scores on all of the tests. It is found by dividing the range on each test (Col. 5) by the sum of the ranges (Total, Col. 5), expressing the result as a percentage. (Extend the result to the third decimal, round off to two places, and drop the decimal point.) The sum of these "contributions" should be approximately 100; rounding off the percentage may account for a slight error.

h. Column 7, Theoretical Multiplier.

This number indicates what "apparent" weight should be given to the scores on each test in order to yield the effective contributions that have been set up as desirable in column 2. It is found by dividing the *Desired Contribution* (Col. 2) by the *Contribution if Added* (Col. 6). The division should be carried to the second decimal place and rounded to one decimal place. If the two "contribution" values are equal, the result will be 1. If the Desired Contribution is larger than the percentage recorded in column 6, the resulting theoretical multiplier will, of course, be greater than 1; if the Desired Contribution is smaller, obviously the theoretical multiplier must be less than 1. Check the theoretical multipliers to be sure that they follow this rule. *The number in column 2 must always be divided by the number in column 6.* (Another check can be made. Multiply each range by its

theoretical multiplier and divide the product by the desired contribution of that test. If the computations have been correctly made, the resulting quotients should be approximately equal.)

i. Column 8, Practical Multiplier.

The theoretical multipliers found in column 7 may not be suitable for easy and rapid use. Since the original judgment as to how much each test should contribute was an approximate one, it is permissible to round off the multipliers to figures that will be convenient and simplify the application of the weights. This may be done either by directly rounding the numbers in column 7, or by finding numbers that will be roughly proportional to them. In the example given in the table the theoretical multipliers were rounded to small whole numbers or simple fractions with the exception of that for Test D which was left at 1½, a number that is fairly easy to use as a multiplier. The desired contributions would have been more nearly obtained, however, if the following "Practical Multipliers" had been used: Test A, 1½; Test B, 1; Test C, 4 or 5; Test D, 2 or 3, Test E, 3; Test F, ½. In deciding on practical multipliers, two additional points should be considered:

(1) *Does the "rounding" increase or decrease the contribution of the test?* Note that in the illustration all of the theoretical multipliers were rounded *upward*. As a result, the contributions of tests B and F were significantly increased, while the contribution of Test D which was not rounded was consequently reduced.

(2) *Does the increase or decrease in contribution caused by the rounding work in the direction desired?* For example, rounding the theoretical multiplier of Test B from 0.7 to 1, makes its contribution almost equal to that of Test A, and rounding the theoretical multiplier of Test F from 0.3 to 0.5 or ½ makes its contribution slightly more than one third greater than that of Test E. Perhaps it would have been preferable, leaving the others as given in the table, to have chosen ½ as the practical multiplier for Test B and 1/3 as the practical multiplier for Test F.

j. Applying the Weights.

Having completed the selection of practical multipliers for each of the tests, the next steps in obtaining the weighted grades for a group are (1) applying the proper multiplier to each man's score on each of the tests, (2) finding the total of the weighted scores, and (3) translating the total weighted scores into grades or marks.

Figure 21, below, will illustrate a method of handling the job of applying the multipliers to each man's scores.

CLASS RECORD SHEET—TEST SCORES, FINAL MULTIPLE, AND GRADE

NAME	*Scores on*	*Test A*	*Test B*	*Test C*	*Test D*	*Test E*	*Test F*	*TOTAL SCORE*	*GRADE*
	Multiplier	1	1	3	1½	2	½		
Allen, James C.	Test Score	82	21	90	55	77	68		
	Weighted	82	21	270	82	154	34	643	
Bacon, Arthur L.	T.S.	70	28	78	42	71	102		
	W.S.	70	28	234	63	142	51	588	
Brighton, Richard	T.S.	65	35	85	61	81	49		
	W.S.	65	35	255	92	162	24	633	
Carter, Lloyd	T.S.	63	15	95	64	82	65		
	W.S.	63	15	285	96	164	32	655	
Crawford, Henry	T.S.	72	14	75	55	76	116		
	W.S.	72	14	225	82	152	58	603	

Figure 21.—Sample class record sheet showing application of weights to test scores.

The following details are worth noting:

(1) In setting up the record sheet, space is allowed in the column headings to insert the multipliers.

(2) Two lines are allowed for each man's record. The raw scores on the tests are entered on the first line. This can be done as soon as the papers are scored. The second line is reserved for the weighted score, which can not be determined until after all of the tests have been given and the multipliers decided upon. Having the weighted scores on a separate line simplifies computation of the Total Score as there is less danger of picking up a raw score instead of a weighted score than there would be if the man's record were kept on a single line with spare columns for the weighted scores. If a record sheet is arranged with spare columns for the weighted scores it is suggested that they be written in with a different color ink.

(3) There is less chance of error in computing the weighted scores if the weight (multiplier) is applied to all of the individual scores of one test, then the determined weight applied to all individual scores of the next test, etc., rather than applying all of the weights to the test scores for one man at a time. In other words, when figuring the weighted scores, work down the columns rather than across the rows.

(4) After the weighted scores on each of the tests have been recorded, add up the numbers in the weighted score line for each man and write the

sum in the TOTAL SCORE column. (There is no need to record the total of the raw scores.)

k. Convert the Total Weighted Scores to Grades.
When the total weighted scores have been computed, they can be translated to grades or marks in the same way as scores on a single test, using a translation graph and conversion table as explained in Chapter V, Section 5C.

7C. SUGGESTIONS ON THE USE OF A WEIGHTING SYSTEM

Here are a few additional pointers on setting up and using a weighting system. Remember that the whole purpose of the weighting system is to help you to give grades that will be fair to the men and fair to the Navy. You don't want to give a 4.0 to a diesel mechanic who knows all there is in the book about the engine but doesn't know how to operate one and keep it in good running condition. The weighting system should work for you, you're not supposed to be working for it.

1. Don't Get Fouled Up on Test Scores That "Run Backwards"

The first illustration of a performance test in Chapter II, Section 2E, describes a performance test for Gunner's Mates in which the *best* performance earns the *lowest* score. If scores in this test are to be included in any combination with tests where high scores mean good performance, the raw scores must be *inverted*. This can be done either before or after the multipliers for weighting have been decided, but it *must* be done before the TOTAL SCORE is found. To *invert* the scores, simply take any number greater than the "highest" score on the test and subtract each score from it. For example, if the scores on a test range from 12 to 27, you might use 30 as the constant number from which to subtract each score. The *inverted* scores would then range from 18 to 3.

2. Any Weighting System Should Be Simple

In general the multipliers should be whole numbers or fractions that are easily used. In some cases where such precision seems justified, more complex multipliers (like 3/4, 5/3) may be used but such situations are not common in the average training program. Since it is the *proportions* that count, rather than the absolute size of the multipliers, it may be simpler to *divide* the scores on one or two tests rather than to multiply those on several others.

(A multiplier of ½ means that the scores are divided by 2.) For example, if the multipliers on a series of tests work out to be 1, 2, 2, and 4, it is obvious that the same relative weights will be obtained with less work if the multipliers used are ½, 1, 1, and 2.

3. No Large Multipliers Should Be Necessary

In general, if the range on a test is so small that large multipliers seem necessary, something should be done to improve the test. It may be necessary to use the large multiplier with the test scores already obtained, but the small range is a danger signal indicating that the test may be inaccurate and unreliable and therefore

should not be given as much importance in the final grade as was originally placed in the "Desired Contribution" column. To increase the actual range of the test:

a. Make the test longer, by adding more items or more sub-tests. This step may or may not be effective if the added items are very easy or very difficult.
b. Look for the items which are too easy (almost everyone gets the right answer) and those which are too hard (almost no one gets the right answer) and replace these items with others that are of more nearly average difficulty.

4. Revision of the Scoring System on a Test May Reduce the Task of Weighting

Often the scoring system used for a test allows too many or too few points per item. If the scoring system is changed, the range may be altered so that no multiplier need be used. For example the points allowed for a "time bonus" for quality of product on a performance test may be changed, or the scoring plan for identification tests may require correct selection of *both* part name and part function instead of giving separate credit for each.

5. Convert Scores to Grades or Marks Only After the Total Scores Have Been Obtained

If the system of weighting test scores described in this chapter is followed there is no necessity for converting scores on a single test to grades. One conversion can be made after the weighted scores have been added. However, it should be noted that translating the scores on a test to grades, *as directed in Section 5C* will have the effect of *making the range of grades* on the tests approximately equal. If you are working with *grades* instead of *scores,* the multipliers can be made directly proportional to the "Desired Contributions" of the measures. This applies, however, only if the grades are distributed over the whole range of the grading system from slightly below passing to nearly the top grade. If a system of conversion to grades has been used which does not meet this condition, then grades should be weighted in the same manner as has been outlined for scores in this chapter. If both test scores and recitation grades are used in determining the final grades, grades on the 63-99 scale can be handled in exactly the same way as test scores, but marks on the 4.0 scale should be handled by moving the decimal point one place to the right.

6. Make a Final Conversion Even if the Elements of the Combination Are Grades

As was pointed out in connection with section 7A3, when several measures are combined, the sum of the highest scores will be greater than the highest total score made by one individual unless that person had the highest score on all of the measures. If the final grades are computed simply by dividing each man's total weighted score by the sum of the multipliers (that is by finding the weighted average) the range of the final grades will almost always turn out to be smaller than the range of most of the grades included in it. Further, the greater the number of tests included in the average the greater this reduction in range is likely to be.

7. Recheck the Multipliers

It may seem that the same series of tests given to two different classes, perhaps six months apart, should be given the same weights both times. This may prove true, but it can not be assumed. The range on one or more of the tests may vary enough to require a change in the multipliers. The addition of a new test to the series, or the revision of a test previously used, will of course require a new determination of the ranges and relative weights to be assigned to all of the tests used.

USING THE EDGE-MARKED ANSWER SHEET

(Answer Sheet Test Form, NavPers 1550/2)

NOTE: This supplement to Constructing and Using Achievement Tests, NavPers 16808-B adds to the guidance given in Chapter V, "Scoring Tests and Grading Students," and Chapter VI, "Interpretation of Test Results." There are also implications for change in emphasis at other points in the manual.

1. Applying Test Results to Improve Instruction.

Testing can be used as a tool for guiding instruction. For such use the instructor is more concerned with the answers to particular questions, or sets of questions, than with the scores made, though the scores are important too in finding out which trainees are ready to go ahead and which ones need more help. The information he can get from the trainees' answers will help him to adapt his lesson plans and procedures to their actual progress and needs, rather than to some achievement that is assumed to have resulted from past experience and learning. If the information is "fed back" to the class, either as a group or in sub-groups or individually, the trainees themselves can get a better sense of their progress and of what they need to do in specific areas of weakness or failure.

The earlier this "feedback" occurs the more effective it will be. This is true both in terms of motivation and in terms of relating the information to specific learning involved. Trying to connect success or particular failure/error with choices made three days ago isn't a very inspiring or profitable exercise. But immediate recognition of success or error, while the situation is fresh in mind, can encourage and challenge.

The time required to score and analyze the responses to test questions has been a factor in restricting the use of test results for teaching. The processing time can be very significantly reduced by using an edge-marked answer sheet. Some of the feedback, both to the teacher and to the trainees, can be provided immediately following completion of the test, within the same class period.

2. Description.

The edge-marked separate answer sheet is a special form which lends itself to quick visual analysis of the responses given by the students in a class. In other separate answer sheets, assigning each response to a particular location on the page made the scoring of individual papers by punched keys or electrical sensing machines possible. Moving all of the response locations to the edges of the sheet makes them instantly accessible for visual inspection.

Answer Sheet Test Form (NavPers 1550/2) is designed for use with tests made up of four alternative items. It provides spaces for responses to 100 four-choice items (50 on the face, 50 on the back), arranged around all four margins of the page, with the spaces for marking extended to the edge of the sheet. The central portion of the sheet, within the "frame" made by the item-response spaces, may be used to record information requested by the instructor. Labelled spaces for the identifying information usually required appear at the top of this area on the face of the form. Calculations, diagrams, responses to "short answer" (completion type) items or "essay" questions can be given in the remainder of the space.

3. Use in Immediate Class Discussion.

When edge-marked answer sheets are used with a "quiz" on material assigned for out-of-class preparation, or in a review of previously studied topics, the instructor can make an immediate check on the degree of mastery or error on particular concepts or operations sampled in the questions given. This enables him to conduct his critique and remedial instruction immediately, while the trainees have the questions before them, and how they answered is fresh in their minds. Also, he can pinpoint his discussion to commend and reinforce the thinking that led to right choices on significant items, and to explain why other choices that were actually made were not as good, or showed error in knowledge or judgment. Such discussion helps the trainees to learn; is more effective than a delayed explanation of how they got their scores or grades.

The technique for such an on-the-spot analysis of trainee responses is quite simple. The instructor gathers the answer sheets while the students retain the test or "quiz" questions, and arranges them in a stack, face up. He selects particular items or sequences to be checked, and simply by bending and "fanning" the packet he can observe how many answers are "out of line," and which alternatives were most frequently chosen. He may also respond to students' questions about particular items. If the instructor has included critical errors among the responses to certain items on informal tests, he can check the number of students who chose these responses as a guide to the emphasis to be given such material, either in the critique or in later instruction.

4. Marking the Answer Sheets.

Marking is a processing step that prepares the answer sheets for scoring, for "item analysis," or for use in counseling trainees. The same

marking operation can serve all three processes. Basically it consists in comparing the responses chosen or given by individual trainees with those prescribed in a "key" or "gouge" and noting the instances of their being the same, or different.

One way of recording these item-by-item comparisons is to transfer the "key" answers to each of the trainees' answer sheets, using a colored ink or crayon so that it will not be confused with the trainee's response. When the edge-marked answer sheet is used, this can be done easily, using a straight-edge and a felt-tipped marker. The papers are aligned on a flat surface, spaced so as to expose about one-fourth inch (the marked margin) of each. (The heavy line between spaces 3 and 4 on the face and between 62 and 63 on the back will identify papers that are rotated or reversed in the packet.) With a key sheet at the top (or at both top and bottom) of the set of papers, a colored line is drawn across the exposed edges of all the papers, through the response space "keyed" as correct for each item. (A line of a different color may also be used to indicate responses that represent critical error, or for which "penalty" scoring might be applied.) After the 15 items on the left margin have been marked, the papers are realigned to expose the bottom margin, etc. (If there were more than 50 items in the test and the number of papers is small enough, the set can be turned over in order to mark exposed spaces 51-65 on the back left margin before realigning to mark items on the bottom margin.)

The arrangement for marking by transferring the "key" to students' answer sheets is illustrated in Figure 1.

5. Scoring the Papers.

If the trainees' papers have been marked as described above, scoring them is simply a matter of applying whatever rules have been established in the school, or agreed on for the particular test. How the scoring will be done may depend on how the test results are to be used, or on whether the same test will be reused with the same or with other groups of trainees. For some uses, part scores are required and a total score, if computed at all, is only a by-product. If the test results are to be used as a basis for remedial teaching, or for changing the emphasis given or the approach to particular topics, part scores are needed. The conditions under which the test is used may determine whether or not a "correction for guessing" should be applied. It is possible, of course, to rescore the same sets of answer sheets in different ways to meet different requirements.

As a general practice, separate counting of "right," "wrong," and omitted items is recommended. Even if only the "right" responses, without corrections or penalty points, will be used in grading, the separate count is an added assurance that the trainee's reaction to each item has been accounted for, and the repeated comparison of error scores and omissions with the "success" scores of individual students helps the instructor to recognize the relative strengths and weaknesses of his trainees, and of his teaching.

(NOTE: In this illustration, key sheets are used at both top and bottom of the stack. The item numbers are exposed on the bottom key, as an added assurance that the proper edge is being used. A similar arrangement, with or without the key sheet, is used in counting responses made by sample groups for item analysis.)

Figure 1.—Arrangement of students' answer sheets and key for marking.

Related to the application of a correction for guessing, is the question of whether the trainees should be advised or encouraged to guess when they are taking the tests. When the tests are given during a course of instruction, primarily to give the teacher information to use in planning further work with the same students, encouragement of guessing would be contrary to the purpose of testing, as well as contrary to Navy doctrine on responsibility. Guessing, distinguished from choosing the more reasonable of available alternatives, is a form of "bluff" or irresponsible reporting. The trainee should be advised to exercise the same kind of judgment that would be required if the question arose in connection with duty. Some will guess, just as some men will take random, rather than reasoned, action when confronted with an unfamiliar situation or a difficult problem. Confirmation, or correction, of a reasoned choice is likely to have more lasting effect on the trainee's thinking and action than the discovery that a "guess" was "lucky" or otherwise. An effective teaching-learning relationship is more likely to develop if guessing is discouraged not only in the instructions for taking tests, but also in using the "correction" in scoring, and even penalties for critically wrong choices.

Under some circumstances it may be undesirable to transfer the "key" to the answer sheets of all of the students; for example, if substantially the same test is to be used again and the answer sheets are redistributed to the trainees. Even marking items as "right" (or "wrong") could compromise the "key" if the students chose to develop their own "gouge." In these instances another technique of scoring can be used. An expanded scoring "key" can be prepared by cutting off the margins of a keyed answer sheet and remounting them on a standard size sheet so that the numbered response spaces will extend beyond the edges of the standard size sheet. This is essentially the same as arranging four "strip scoring sheets" as a frame in which each trainee's answer sheet can be placed for scoring. (See Figure 2.)

The expanded scoring sheet can be mounted on a small drawing board with slightly raised thumb tacks set just at the edges of the top and one side of the standard size sheet, thus serving to align the papers inserted for comparison scoring.

6. Analyzing Test Results.

The use of test results in improving instruction requires some analysis of the responses given by representative trainees in relation to what had been planned as their learning objectives. Basically, the analysis of test results involves comparisons among the responses chosen by different groups, using the responses agreed on by the "experts" as the common reference. The focal question in planning an analysis is, "For what groups will comparison of responses yield information pertinent to decision or action?" Once the groups to be compared are identified, the technical design for limiting or sampling the papers can be developed.

For application in remedial teaching, and in adapting instruction to the needs of some trainee groups not typical of the usual input, it is more important to know what responses <u>other</u> than the "best" are attractive

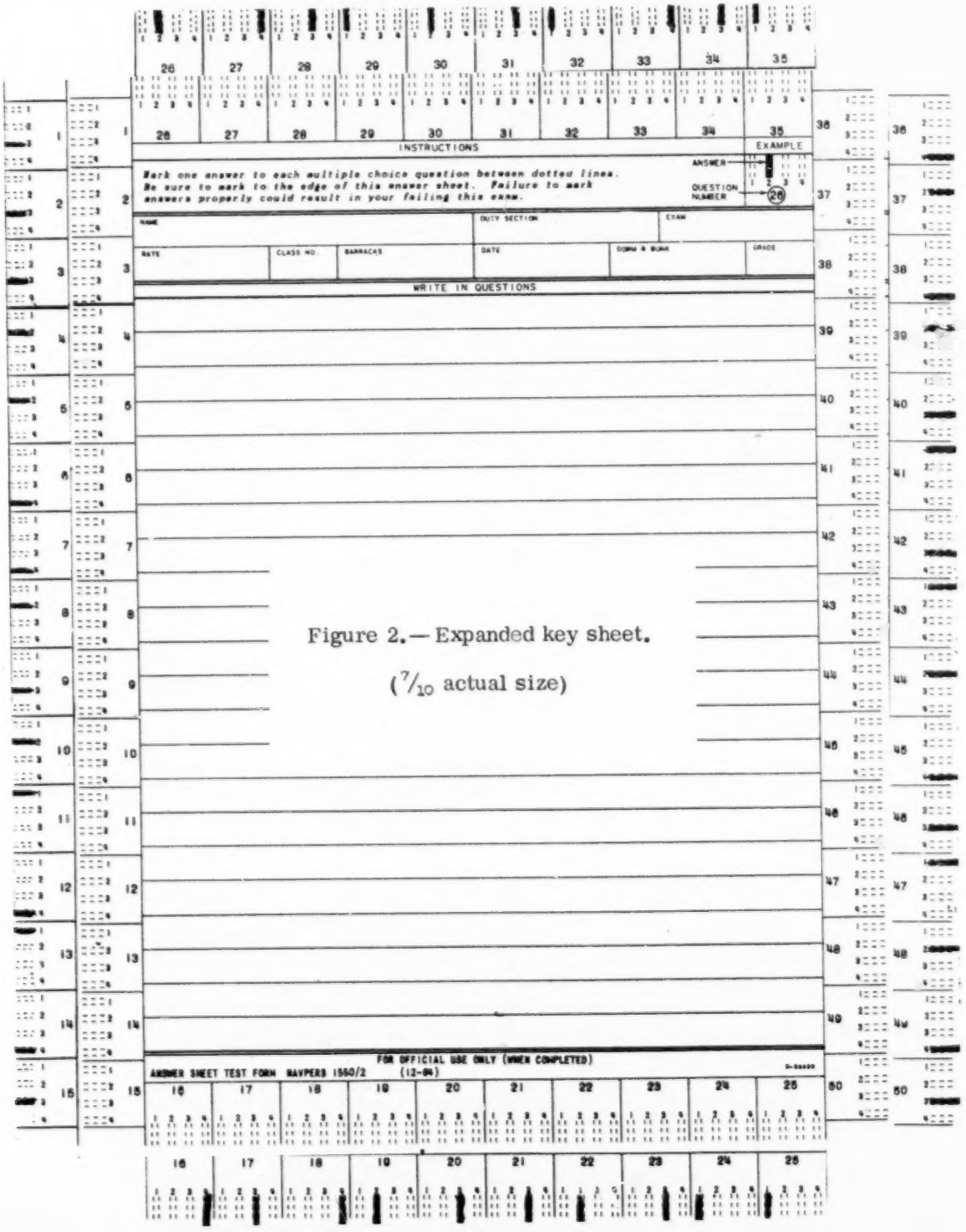


Figure 2.— Expanded key sheet.

($^{7}/_{10}$ actual size)

than to know how many chose the "right" response. The differences in the other responses chosen on some items, as between high-scoring, average or typical, and low scoring trainees may be more significant, practically, than the statistic of their comparative percentage of "right" answers.

Item analysis can be useful in assessing the effects of instruction by comparing the responses made before and after a unit is taught. In such studies, changes in the pattern of errors may be more significant than the gains in scores made.

The effects of different teaching/learning procedures can also be compared. It is possible for comparable groups of trainees, taught in different ways to achieve comparable scores or gains, but with very different patterns of responses to the items comprising the end test. Even when the scores typically achieved by groups taught by different methods are clearly different, the different patterns of response may help to indicate particular features of the methods used which proved effective.

In the usual school situation, where the comparisons needed involve small groups of trainees and the information is intended primarily for local use, the data will ordinarily be "hand processed," rather than set up for machine or electronic processing. Whole groups rather than "representative samples" can be used for analysis; or typical groups, excluding exceptional cases or extremes can be compared.

When the edge-marked answer sheets are used, the tedious and error-prone tallying operation isn't necessary. Instead, the answer sheets are spread in the same manner as for marking, described in Section 4, above. For each group or sub-group, the number of marks in each response space for each item is then counted, and the number recorded on a copy of the test paper, along with a count of the number of papers in which no response was given to an item.

Where exactly the same test was given to two or more groups, the comparisons can be made directly by providing as many columns as there are groups to be compared. Sometimes the same set of items will be used in a different order in different "forms" of a test. The answer sheet for each form must be kept separate, and the count recorded on a copy of the corresponding form. Cross-identification of the items can be made on one form, or the item numbers on the different forms can be entered on an item card to which the data are then transferred. Similarly, some of the same items may be reused in several tests, each of which may include other questions; comparison of item data from such dissimilar tests requires special care in handling and interpretation.

Item analysis for the improvement of the tests themselves can also be made more meaningful with the simplified procedure. The usual procedure of estimating "difficulty" by averaging the percentage of "right" answers to an item for high-scoring and low-scoring groups, and obtaining an index of discrimination based on the difference between these percentages doesn't take into account the responses actually made by the "average" or "middle" members of the group. The assumption that the percentage "right" ("p-value") for the middle group will be somewhere between those of the "high" and "low" groups isn't always borne out, and the

items which prove either easier or harder for the "average" than for both the "high" and "low" groups are particularly likely to need careful study.

The "difficulty" of a test item isn't something inherent in the item itself, but depends on the ability and/or preparation of the group being tested. Obviously the "difficulty" of the same item would be greater for an untaught (pre-training) group than for one which had just completed a course to which the item pertains. The difficulty of items used in "progress" tests is likely to be different (either greater or less) than when the same items are used in final examinations, or in delayed testing several weeks or months after completion of training. Also, the difficulty of an item varies with the "attractiveness" of the alternatives to the persons in the group being tested, especially in the case of true "best answer" questions.

Similarly the "discrimination index" for an item cannot be considered constant. The discrimination index is sometimes referred to as "item validity." It is essentially an estimate of how much the item contributes to placing people in the same order as they would have when ranked according to scores on the whole test, or ratings by instructors or supervisors. Partly, then, the discrimination index for an item depends on its relative difficulty for the group being tested, which was just shown to be variable; partly on the context of other items contributing to the score, or elements considered in the ratings, against which it is measured.

For informal tests used in the schools the question of whether a particular item should be used, modified, or abandoned for application to particular testing situations has to depend on whether it provides useful information about the achievement of the trainee, or his readiness to undertake the next steps of a training program. If it does, the statistics of its indexes of difficulty and of discrimination are "nice to know," but relatively unimportant. If items that seem relevant and appropriate prove extremely difficult or extremely easy, or if their discrimination indexes turn out to be low or negative, they may need to be considered carefully for revision, or for application to a different level of training, or both.

7. Using Test Results in Counseling Trainees.

The information that an instructor gains about the achievement of his trainees, individually through marking and scoring their test papers and collectively through "item analysis" studies, can be turned to good use in his conduct of further teaching and counseling relationships. He should recognize, however, that the test-derived information represents only a small momentary sample of the students' developing patterns of work (behavior). In his teaching and counseling with them, this additional information has to be interpreted in the light of what he has learned in past relationships with them, and can lead to further inquiries about their aims and needs. The test results may furnish the occasion for counseling, but the counseling should be in terms of the trainees' whole pattern of growth and development as related to career and learning objectives, not merely a review of test performance.

The analysis of test results may indicate needs for three different general kinds of remedial or supplementary instruction.

a. Occasionally the teacher will discover that all or most of a class turned up with a very different idea about some topic from what he had intended to teach them, or that even the faster learners are still confused. Normally this should lead to a review lesson (or series) planned with a different approach and techniques. (The teacher may also want to change the way he teaches that topic to later classes of the same kind.)

b. Sometimes inspection of the analysis results will show similar patterns of error over one or more clusters of related items for particular sub-groups in the class. In a sense, the membership of the sub-group is defined by the error pattern, but it may also coincide rather well with some other known grouping of trainees. Some of the scholastically higher ranking members of the class may fall into such a group. When such trainee groups are identified, their special needs can sometimes be met, or more effectively handled, by temporarily making them a sub-section of the class and planning special (different or additional) learning-teaching activities for them. Sometimes such additional/different activities will require organization of "night school" classes, or assignment to special remedial set-back groups.

c. In addition to sub-groups with similar patterns of error, the teacher is most likely to find individuals whose patterns of response are unique. These trainees are likely to need individual counseling by the teacher or a school officer. Such counseling, in turn, might lead to special supplementary or remedial assignments, use of special materials, arrangements for coaching, forming "buddy groups" for mutual assistance in study, etc.

The teacher should also look to the positive side, to identify "success" both by trainee groups and individuals, and in particular aspects of instruction where progress indicates readiness for further development. Discovering or confirming readiness for further learning is the basic purpose for using tests in teaching. To the extent that it is merited, the assurance of readiness found in test results should be passed on to the class. This is more effective if it is given in specific terms, in relation to both the work completed and the new work or material to be introduced. The "formula" for offering such assurance will necessarily vary among the trainees of a class, but even the marginally satisfactory trainees need recognition of what progress they have made if they are to use it effectively in coping with the problem of both "make up" and further "new" learning.

8. Adapting Answer Sheet Test Form (NavPers 1550/2) for Other Item Types.

The lay-out of an answer sheet can be "tailored" to the kinds of test items used, and the order of their arrangement in a particular test. This is practical only if large numbers of students will take the test, or tests with identical distribution of the item types used. Many tests, however, are made up entirely, or nearly so, of items of one type. The most frequently used item type is "multiple choice" or "best answer." In tests developed

in the schools for their own use, the number of alternatives (choices) for such items is usually not more than four. (Inventing additional alternatives is often difficult or illogical, and they are often found ineffective in the sense that very few students choose them.)

The format of Answer Sheet Test Form (NavPers 1550/2) is set up for use with tests in which the items are four-choice; that is with four numbered response spaces in each numbered item space.

There are problems and work situations, of course, which involve relations that are too complex to be represented by test items in which the choices are narrowed down to four. For example, reading and interpretation of diagrams may require cross-relations among symbols, terms, functions, and quantitative values. Such relations might be sampled by matching type test units.

Using the NavPers 1550/2 edge-marked answer sheet does not preclude using such units in a test. With carefully prepared instructions and close supervision, the trainees can modify the lay-out of their answer sheets to adapt them to the requirements of other item types. Usually this would require them to change the numbering of response spaces in specified item spaces, and sometimes might require renumbering of some of the item spaces. Such renumbering must be carefully planned in advance and the answer sheets should be modified and checked before the test papers (questions) are distributed.

In a five-item/eight-response matching unit, each item would require eight response spaces, or the equivalent of two item spaces on the answer sheet. (If the number of responses were from five to eight, it would still be preferable to use two of the item spaces on the answer sheet for each item in the unit; or if 9 to 12, to use three of the item spaces.) Preferably all of the item spaces required for a matching unit should be along one margin; at least the spaces required for any one item should not break around a corner, or from the bottom to the top of the sheet.

As an illustration, assume a test consists of 20 four-choice "best answer" items and two matching units with 5-item/8 response form. Instead of using spaces 21 through 30 for the first matching unit, and 31 through 40 for the second, it would be preferable to use the 10 item spaces in the top margin (26 through 35) for one unit, and 10 of those on the right margin (36 through 45) for the other. If the items are numbered consecutively in the test paper, the item spaces on the answer sheet should be renumbered to conform. Alternatively, the items in the test papers could be numbered to conform with the item spaces assigned to them on the answer sheet. The numbers of the items in the first matching unit would thus become (instead of 21 through 25) 26-7 through 34-5. In either form of adapting, the response numbers in the alternate item spaces (odd numbered, 27 through 45 in this instance) would be changed to 5, 6, 7, and 8.

☆U. S. GOVERNMENT PRINTING OFFICE: 1971—433-578/20